Generalized Hyper Markov Laws For Directed Acyclic Graphs

Emanuel Ben-David
Stanford University

Bala Rajaratnam
Stanford University

Abstract

Recent theoretical work ([15, 9]) in graphical models have introduced classes of flexible multi-parameter Wishart distributions for high dimensional Bayesian inference. A parallel analysis for DAGs or Bayesian networks, arguably one of the most widely used classes of graphical models, is however not available. For Gaussian DAG models the parameter of interest is the Cholesky space of lower triangular matrices with fixed zeros corresponding to the missing arrows of a directed acyclic graph $\mathcal{G}$. In this paper we construct a family of DAG Wishart distributions that form a rich conjugate family of priors with multiple shape parameters for Gaussian DAG models, and proceed to undertake a theoretical analysis of this class with the goal of posterior inference. We first prove that our family of DAG Wishart distributions satisfies the strong directed hyper Markov property. Operating on the Cholesky space we derive closed form expressions for normalizing constants, posterior moments, Laplace transforms and posterior modes, and demonstrate the use of the DAG Wishart class in posterior analysis. We then consider submanifolds of the cone of positive definite matrices that correspond to covariance and concentration matrices of Gaussian DAG models. In general these spaces are curved manifolds and thus the DAG Wisharts have no density w.r.t Lebesgue measure. Hence tools for posterior inference on these spaces are not immediately available. We tackle the problem in three parts, with each part building on the previous one, until a complete solution is available for ALL DAGs.

In Part I we note that when $\mathcal{G}$ is perfect, associated covariance and concentration spaces are open cones and hence we proceed to derive the induced DAG Wishart distribution on these cones. A comprehensive analysis is however only possible for the class of perfect DAGs. In Part II we formally establish that for any non-perfect DAG, covariance and concentration spaces have Lebesgue measure zero in any Euclidean vector space containing them, and hence the DAG Wishart family introduced above does not have a density w.r.t. Lebesgue measure for $\mathcal{G}$ non-perfect. We therefore propose a unified approach for all Gaussian DAG models by appealing to the theory of Hausdorff measure theory. First we derive the functional form of the DAG Wishart density w.r.t Hausdorff measure. We demonstrate however that even for the simplest of graphs, the Hausdorff density is not amenable to posterior analysis. In part III we define new spaces that are projections of covariance and concentration DAG spaces onto Euclidean space that yield natural isomorphisms. We exploit this bijection to derive the densities of DAG Wishart and DAG inverse Wishart distributions w.r.t Lebesgue

measure, and thus avoid recourse to Hausdorff densities. We demonstrate that this third approach is extremely beneficial and is readily amenable for high dimensional posterior analysis. We derive hyper Markov properties and posterior moments for DAG Wishart and inverse Wishart distributions corresponding to arbitrary DAGs, and not just for the class of perfect DAGs.

1 Introduction

Graphical models yield compact representations of the joint probability distribution of a multivariate random vector, and they have therefore proved to be very useful in discovering structure, especially in high-dimensional data. These models use nodes of graphs to represent components of a random vector, and edges between these nodes, to capture the relationships between these variables. In general these graphs can have three types of edges: directed, undirected or bi-directed. Undirected graphs are often used to represent association through conditional independences whereas bi-directed graphs are often used to represent marginal independences. Directed acyclic graphical models (DAGs), or sometimes referred to as Bayesian networks[1], are often used to represent causal relationships among random variables. Graphical Markov models corresponding to DAGs have useful statistical properties, especially in high dimensional settings. The joint probability density function (pdf) of a DAG model factorizes according to the graph into the product of the conditional pdfs for each variable given its parents, and can thus lead to a substantial reduction in the dimensionality of the parameter space. Directed acyclic graphical models have also found widespread use in the biomedical sciences, social sciences and in computer science. Estimating the covariance or inverse covariance corresponding to such DAGs is therefore an important area of research, especially in high dimensional settings.

From a theoretical statistics perspective DAG models correspond to curved exponential families, and are distinctly different from the standard undirected or concentration graph models, which correspond to natural exponential families. Unlike in the natural exponential family setup, a default prior, as given by the Diaconis-Yvilsaker (DY) framework, is not available for a general DAG (see [9] for a thorough discussion). A connection between DAGs and undirected or concentration graphs can however be used to derive a default prior for a subclass called perfect graphs. Indeed if the graph is "perfect"[2], such DAGs are said to be Markov equivalent to decomposable concentration graph models, i.e., they both capture the same set of conditional independences. This connection can be exploited in the sense that inferential tools for perfect DAGs can be borrowed from the decomposable concentration graph setting. More specifically, in their pioneering work, Dawid and Lauritzen [7] developed the DY prior for this class of models when the graph is decomposable. In particular, they introduced the hyper-inverse Wishart as the DY conjugate prior for concentration graph models. This work has been extended by the recent methodological contributions by Letac and Massam [15] who develop a rich family of multi-parameter conjugate priors that subsumes the DY class. Both the hyper inverse Wishart priors and the "Letac-Massam"

[1] Sometimes also called recursive Markov models.

[2] A concept that will be formally defined later.

priors have attractive properties which enable Bayesian inference, with the latter allowing multiple shape parameters and hence suitable in high-dimensional settings. Bayesian procedures corresponding to these Letac-Massam priors have been derived in a decision theoretic framework in the recent work of Rajaratnam et al. [19]. A parallel theory for the class of Gaussian covariance graph models, graphical models which encode marginal independencies, has been recently developed by Khare and Rajaratnam ([9], [10]). With a few exceptions (see [21] for instance), all of the above methodological contributions are for decomposable graph models. Furthermore, the results for undirected or concentration graph models cannot be carried over to DAGs when $\mathcal{G}$ is no longer perfect. This is because the Markov equivalence property between DAGs and undirected graphical models breaks down in the sense that the DAG model, and the undirected graphical model, capture different set of conditional independencies when G is not perfect (or equivalently non-decomposable).

The literature on graphical models in general, and DAGs in particular, is extensive and thus we do not undertake a literature review of the work in this area. We shall however briefly review work that is directly relevant to this paper when notation and terminology is introduced.

The principal objective of this paper is to develop a framework for flexible high dimensional Bayesian inference for Gaussian DAG models, i.e., for the class of Gaussian distributions that have the directed Markov property. The previously established classes of generalized multi-parameter Wishart distributions developed by Letac and Massam [15] in the concentration graph setting, and by Khare and Rajaratnam ([9] in the covariance graph model setting, are not directly applicable to the general DAG setting, though they provide useful insights, as will be demonstrated in this paper.

For Gaussian DAG models the parameter of interest, denoted by $\Theta_{\mathcal{G}}$, is the space of lower triangular matrices with fixed zeros corresponding to the missing arrows of a directed acyclic graph $\mathcal{G}$. We introduce a rich class of generalized multi-parameter DAG Wishart distributions on $\Theta_{\mathcal{G}}$ was proposed, and studied with the explicit goal of Bayesian inference in high dimensional settings. This family extends the classical Wishart distribution in the sense as the latter becomes a special case of our family of DAG Wishart distributions. A comprehensive analysis of this family of generalized Wishart distributions was possible for arbitrary DAGs when working on $\Theta_{\mathcal{G}}$. Indeed analytic expressions for posterior moments, Laplace transforms, posterior modes and hyper Markov properties are established. The hyper Markov property in turn enables the explicit computation of expected values and Laplace transforms in the Cholesky parameterization. Unlike their concentration and covariance graph counterparts, we show that sampling from our Wishart distribution for an arbitrary DAG model does not require recourse to MCMC. Once more we note that for concentration graph models, sampling from the posterior can be done in closed form only for decomposable models. For covariance graph model, sampling from the posterior without recourse to MCMC can be done only for homogeneous graphs. We also show that our DAG Wishart distributions can be derived in an equivalent way using the general approach in [7] under the so-called global independence assumption. The latter approach does not however immediately give a means to specify hyper-parameters that will correspond to our DAG Wishart distributions. We also provide a discussion of the fact that our DAG Wishart distributions are in general different from the Letac-Massam priors. However, when the

underlying graph $\mathcal{G}$ is homogeneous, the Letac-Massam $W_{P_\mathcal{G}}$ priors are a special case of our distributions. We also provide a discussion of the fact that our DAG Wishart distributions are in general different from the priors introduced in Khare and Rajaratnam([9] for covariance graph models.

After introducing our class of flexible Hyper Markov laws we explicitly tackle the question of deriving tools for Bayesian inference for Gaussian DAG models. In order to do this we then consider submanifolds of the cone of positive definite matrices that correspond to covariance and concentration matrices of Gaussian DAG models. In general these spaces are curved manifolds and thus the DAG Wisharts have no density w.r.t Lebesgue measure. Hence tools for posterior inference on these spaces are not immediately available. We proceed to tackle this problem in three parts, with each part building on the previous one, until a complete solution is available for ALL DAGs.

In Part I we derive DAG Wishart densities for perfect DAGs. It was noted above that the space of covariance and concentration matrices corresponding to Gaussian DAGs are in general curved sub-manifolds of Euclidean space. When $\mathcal{G}$ is perfect however, these are open cones and the induced DAG Wishart density on these cones can be derived. We then proceed to derive Laplace transforms and expected values in this setting. Computation of expected values of covariance and concentration matrices corresponding to DAG models is no longer possible with this approach, except when $\mathcal{G}$ is perfect, as the space on which these matrices live are in general curved manifolds. We note that a comprehensive framework for Bayesian inference that goes beyond "perfect" graphs is however critical for practical applications. The induced Wishart and inverse Wishart distributions on concentration and covariance spaces for general DAGs require more sophisticated tools and is the subject of Parts II and III.

In Parts II we undertake the endeavor of deriving the induced Wishart and inverse Wishart densities on covariance and concentration spaces corresponding to arbitrary DAGs. We first establish that for any non-perfect DAG, covariance and concentration spaces have Lebesgue measure zero in any Euclidean vector space containing it, and hence the DAG Wishart family $\pi_{U,\alpha}^{P_\mathcal{G}}$ introduced in our previous work does not have a density w.r.t. Lebesgue measure. We propose to overcome this in two novel ways. First we derive the functional form of the density of $\pi_{U,\alpha}^{P_\mathcal{G}}$ w.r.t Hausdorff measure by developing the appropriate tools which allow us to work on concentration spaces corresponding to DAGs. This approach entails working with curved manifolds and Hausdorff measures on arbitrary metric spaces. We then proceed to demonstrate that even for the simplest of graphs, the Hausdorff density is not amenable to posterior analysis.

In Part III we define new spaces that are projections of covariance and concentration DAG spaces onto Euclidean space that yield natural isomorphisms. In particular, these new spaces termed as the space of "incomplete" covariance and concentration spaces correspond to functionally independent elements of the covariance and concentration matrix of Gaussian DAG models. Given incomplete matrices from these spaces, it is always possible to "complete" them in polynomial time, so that the completion corresponds to covariance and concentration matrices of Gaussian DAG models. We exploit these bijections to derive the densities of DAG Wishart and DAG inverse Wishart distributions w.r.t. Lebesgue measure and thus avoid recourse to Hausdorff densities. We demonstrate that the latter

approach is novel, extremely beneficial, and is readily amenable for high dimensional posterior analysis. We then proceed to establish hyper Markov properties and derive posterior moments for DAG Wishart and inverse Wishart distributions corresponding to arbitrary DAG models and not just for the class of perfect DAGs. In doing so we succeed in developing a unified framework for all Gaussian DAG models - that is suitable for both perfect and non-perfect DAGs. Our approach also allows us to formally demonstrate that the class of inverse DAG Wisharts introduced in this paper naturally contains an important sub-class of inverse Wishart distributions for that was introduced by Khare and Rajaratnam [9] in the context of Gaussian covariance graph models.

Table 1 summarizes the properties of the various mulit-parameter Wishart distributions that have been recently introduced to the mathematical statistics literature for use in Gaussian graphical models. It is clear from this table that the Wishart distributions introduced in this paper are applicable in all generality - and not just when the graph is perfect, or equivalently, decomposable, and in this sense very powerful. The ability to specify the induced Wishart distributions and posterior moments for arbitrary graphs is especially useful.

	DAG			UG			COVG		
	ALL	P	H	ND	D	H	ND	D	H
Conjugacy property	✔	✔	✔	✗	✔	✔	✗	✔	✔
Normalizing constant in closed form	✔	✔	✔	✗	✔	✔	✗	✗	✔
Posterior moments in closed from	✔	✔	✔	✗	✔	✔	✗	✗	✔
Posterior mode in closed from	✔	✔	✔	✗	✔	✔	✗	✗	✔
Hyper Markov properties	✔	✔	✔	✗	✔	✔	✗	✗	✔
Tractable sampling from the distribution	✔	✔	✔	✔	✔	✔	✗	✔	✔

Table 1: Properties of Wishart distributions for the three classes of Gaussian graphical models.
Abbreviations. ND: Non-decomposable, D/P: Decomposable/Perfect, H: Homogeneous.

This paper is structured as follows. Section 2 introduces required preliminaries, notation and Section 3 formally defines Gaussian DAG models and parameterization corresponding to Gaussian DAG models. The introduction, preliminaries and parameterizations for DAG models are discussed in some detail to make the paper self-contained and for establishing consistent notation. These sections can be skipped by a reader familiar with the subject matter. In Section 5, the class of generalized Wishart distributions for Gaussian DAG models are formally constructed. Conjugacy to the class of Gaussian DAG models and necessary and sufficient conditions for integrability are established. Furthermore,

comparison to conjugate priors in concentration graph and covariance graph models is undertaken. Section 6 establishes hyper Markov properties for our family of priors. In Section 7 we evaluate Laplace transforms, posterior moments and posterior modes for our class of distributions corresponding to the Cholesky parameterization when G is an arbitrary DAG.

Analysis of our DAG Wishart distributions on corresponding covariance and concentration spaces with a view to developing tools for high dimensional Bayesian inference using the class of DAG Wishart distributions in three Parts. Part I considers the class of perfect DAGs and derives posterior quantities. Part I (Section 8) derives the induced DAG Wishart densities on covariance and concentration spaces for perfect DAGs. We the proceed to show that the expected values of the covariance and concentration matrix can be computed easily for perfect DAGs.

Part II (Section 9) introduces derives the density of our priors w.r.t. Hausdorff measure when $\mathcal{G}$ is arbitrary, i.e., no longer perfect. Part III (Sections 10 and Section 11) defines functionally independent projection of spaces of concentration and covariance matrices that correspond to arbitrary DAG models and proceed to derive the induced measure of our class of Wishart distributions on these spaces. Consequently we proceed to establish hyper Markov properties and derive the expected value of the covariance and concentration matrix for DAG models. We also demonstrate that when $\mathcal{G}$ is no longer perfect the class of DAG Wishart distributions do not belong to the class of general exponential families. Section 12 concludes by summarizing the results in the paper.

2 Preliminaries

In this section, we give the necessary notation, background and preliminaries required in subsequent sections.

2.1 Graph theoretic notation and terminology

In this subsection, we introduce some necessary graph theoretic notation and terminology. Our notation presented here closely follows the notation established in [13], [6].

A graph $\mathcal{G}$ is a pair of objects (V, E), where V is a finite set representing the vertices (or nodes) of $\mathcal{G}$; and E is a subset of $V \times V$ consisting of the edges. An edge $(i, j) \in E$ is called directed if $(j, i) \notin E$. We write this as $i \longrightarrow j$ and say that i is a parent of j, and that j is a child of i. The set of parents of a vertex j is denoted by $pa(j)$, and the set of children of a vertex i is denoted by $ch(i)$. The family of j, denoted by $fa(j)$, is $fa(j) = pa(j) \cup \{j\}$. Two distinct vertices i and j are said to be adjacent if (i, j) or (j, i) are in E, i.e., if there is any type of edge, directed or undirected, between these two vertices. We write $i \sim j$ if there is an undirected edge[3] between vertices i and j and say that i is a neighbor of j, j is a neighbor of i, or i and j are neighbors. The set of neighbors of i is denoted by $ne(i)$.

More generally, for $A \subset V$ we define $pa(A)$, $ch(A)$, $ne(A)$ and $bd(A)$ as the collection

[3]Note that in enumerating the number of edges of a graph, each undirected edge, though consisting of two pairs, counts only once.

of the parents, children, neighbors, and boundary respectively, of the members of A, but excluding any vertex in A:

$$pa(A) = \cup_{i\in A} pa(i) \setminus A, \quad ch(A) = \cup_{i\in A} ch(i) \setminus A, \quad ne(A) = \cup_{i\in A} ne(i) \setminus A,$$

An undirected graph, "UG", is a graph with all of its edges undirected, whereas a directed graph, "DG", is a graph with all of its edges directed. We shall use the symbol $\mathcal{G}$ to denote a general graph, and make clear within the context in which it is used, whether $\mathcal{G}$ is undirected or directed.

We say that the graph $\mathcal{G}' = (V', E')$ is a subgraph of $\mathcal{G} = (V, E)$, denoted by $\mathcal{G}' \subset \mathcal{G}$, if $V' \subset V$ and $E' \subset E$. In addition, if $\mathcal{G}' \subset \mathcal{G}$ and $E' = V' \times V' \cap E$, we say that $\mathcal{G}'$ is an induced subgraph of $\mathcal{G}$. We shall consider only induced subgraphs in what follows. For a subset $A \subset V$, the induced subgraph $\mathcal{G}_A = (A, A \times A \cap E)$ is said to be the graph induced by A. A graph $\mathcal{G}$ is called complete if every pair of vertices are adjacent. A clique of $\mathcal{G}$ is an induced complete subgraph of $\mathcal{G}$ that is not a subset of any other induced complete subgraphs of $\mathcal{G}$. More simply, a subset $A \subset V$ is called a clique if the induced subgraph $\mathcal{G}_A$ is a clique of $\mathcal{G}$.

A path of length $k \geq 1$ from vertex i to j is a finite sequence of distinct vertices $v_0 = i, \ldots, v_k = j$ in V and edges $(v_0, v_1), \ldots, (v_{k-1}, v_k) \in E$. We say that the path is directed if at least one of the edges is directed. We say i leads to j, denoted by $i \longmapsto j$, if there is a directed path from i to j. A graph $\mathcal{G} = (V, E)$ is called connected if for any pair of distinct vertices $i, j \in V$ there exists a path between them. An n-cycle in $\mathcal{G}$ is a path of length n with the additional requirement that the end points are identical. A directed n-cycle is defined accordingly. A graph is acyclic if it does not have any cycles. An acyclic directed graph, denoted by DAG (or ADG), is a directed graph with no cycles of length greater than 1.

The undirected version of a graph $\mathcal{G} = (V, E)$, denoted by $\mathcal{G}^{\mathrm{u}} = (V, E^{\mathrm{u}})$, is the undirected graph obtained by replacing all the directed edges of $\mathcal{G}$ by undirected ones. An immorality in a directed graph $\mathcal{G}$ is an induced subgraph of the from $i \longrightarrow k \longleftarrow j$. Moralizing an immorality entails adding an undirected edge between the pair of parents that have the same children. Then the moral graph of $\mathcal{G}$, denoted by $\mathcal{G}^{\mathrm{m}} = (V, E^{\mathrm{m}})$, is the undirected graph obtained by first moralizing each immorality of $\mathcal{G}$ and then making the undirected version of the resulting graph. Naturally there are DAGs which have no immoralities and this leads to the following definition.

Definition 2.1. A DAG $\mathcal{G}$ is said to be "perfect" if it has no immoralities; i.e., the parents of all vertices are adjacent, or equivalently if the set of parents of each vertex induces a complete subgraph of $\mathcal{G}$.

Given a directed acyclic graph (DAG), the set of ancestors of a vertex j, denoted by $an(j)$, is the set of those vertices i such that $i \longmapsto j$. Similarly, the set of descendants of a vertex i, denoted by $de(i)$, is the set of those vertices j such that $i \longmapsto j$. The set of non-descendants of i is $nd(i) = V \setminus (de(i) \cup \{i\})$. A set $A \subset V$ is called ancestral when A contains the parents of its members. The smallest ancestral set containing the subset B of V is denoted by $An(B)$.

2.2 Decomposable graphs

An undirected graph $\mathcal{G}$ is said to be decomposable if no induced subgraph contains a cycle of length greater than or equal to four. The reader is referred to Lauritzen [13] for all the common notions of decomposable graphs that we will use here. One such important notion is that of a perfect order of the cliques. Every decomposable graph admits a perfect order of its cliques. Let $(C_1, \cdots, C_k)$ be one such perfect order of the cliques of the graph $\mathcal{G}$. The history for the graph is given by $H_1 = C_1$ and

$$H_j = C_1 \cup C_2 \cup \cdots \cup C_j, \quad j = 2, 3, \cdots, k,$$

and the (minimal vertex) separators of the graph are given by

$$S_j = H_{j-1} \cap C_j, \quad j = 2, 3, \cdots, k.$$

Let

$$R_j = C_j \setminus H_{j-1} \text{ for } j = 2, 3, \cdots, k.$$

Let $k' \leq k-1$ denote the number of distinct separators and $\nu(S)$ denote the multiplicity of S, i.e., the number of j such that $S_j = S$. Generally, we will denote by $\mathscr{C}_\mathcal{G}$ the set of cliques of a graph and by $\mathscr{S}_\mathcal{G}$ its set of separators.

2.3 Markov properties for directed acyclic graphs

Let V be a finite set of indices and $(X_i)_{i\in V}$ a collection of random variables, where each X_i is a random variable on the probability space $\mathcal{X}_i$. Let the probability space $\mathcal{X}$ be defined as the product space $\mathcal{X} = \times_{i\in V}\mathcal{X}_i$. Now let $\mathcal{G} = (V, E)$ be a DAG. For simplicity, and without loss of generality, we always assume that the given DAG $\mathcal{G}$ is connected and the edge set E contains all the loops $(i, i), i \in V$[4]. We say that a probability distribution P on $\mathcal{X}$ has the recursive factorization property w.r.t. $\mathcal{G}$, denoted by DF (the directed factorization property), if there are σ-finite measures μ_i on $\mathcal{X}_i$ and nonnegative functions $k^i(x_i, x_{pa(i)})$, referred to as kernels, defined on $\mathcal{X}_{fa(i)}$ such that

$$\int k^i(y_i, x_{pa(i)}) d\mu_i(y_i) = 1, \quad \forall i \in V,$$

and P has a density p, w.r.t. the product measure $\mu = \otimes_{i\in V}\mu_i$, given by

$$p(x) = \prod_{i\in V} k^i(x_i, x_{pa(i)}).$$

In this case, each kernel $k^i(x_i, x_{pa(i)})$ is in fact a version of $p(x_i|x_{pa(i)})$, the conditional distribution of X_i given $X_{pa(i)}$. An immediate consequence of this definition is the following lemma.

[4]For convenience we draw the graphs without their loops.

Lemma 2.1. *(from [13]) If P admits a recursive factorization w.r.t. the directed graph $\mathcal{G}$, then it also admits a factorization w.r.t. the undirected graph $\mathcal{G}^{\mathrm{m}}$, and, consequently, obeys the global Markov property*[5] *w.r.t. $\mathcal{G}^{\mathrm{m}}$.*

Proof. Note that for each vertex $i \in V$ the set $fa(i)$ is a complete subset of $\mathcal{G}^{\mathrm{m}}$. Thus if we define $\psi_{fa(i)}(x_{fa(i)}) = k^i(x_i, x_{pa(i)})$, then $p(x) = \prod_{i\in V} p(x_i|x_{pa(i)}) = \prod_{i\in V} k^i(x_i, x_{pa(i)}) = \prod_{i\in V} \psi_{fa(i)}(x_{fa(i)})$. Therefore, P admits a factorization w.r.t. $\mathcal{G}^{\mathrm{m}}$ and by proposition 3.8 in [13] it also obeys the global Markov property w.r.t. $\mathcal{G}^{\mathrm{m}}$. □

Another direct implication of the DF property is that if P admits a recursive factorization w.r.t. $\mathcal{G}$, then, for each ancestral set A, the marginal distribution P_A admits a recursive factorization w.r.t. the induced graph $\mathcal{G}_A$. Combining this result with Lemma 2.1 we obtain the following: P admits a recursive factorization w.r.t. $\mathcal{G}$ then $A \perp B|S$ $[P]$, whenever A and B are separated by S in $(\mathcal{G}_{An(A\cup B\cup S)})^{\mathrm{m}}$. We call this property the directed global Markov property, DG, and any distribution that satisfies this property is said to be a directed Markov field over $\mathcal{G}$. For DAGs the directed Markov property plays the same role as the global Markov property does for undirected graphs, in the sense that it provides an optimal rule for recovering the conditional independence relations encoded by the directed graph.

We now introduce below another Markov property for DAGs. A distribution P on $\mathcal{X}$ is said to obey the directed local Markov property (DL) w.r.t. $\mathcal{G}$ if for each $i \in V$

$$i \perp nd(i)|pa(i).$$

Now for a given DAG $\mathcal{G}$ consider the so-called "parent graph" defined as follows: The parent graph $\mathcal{G}_{par}$ of $\mathcal{G}$ is a DAG isomorphic to $\mathcal{G}$ and obtained by relabeling the vertex set V as $1, 2, \ldots, |V|$, in such a way that $pa(i) \subset \{i+1, \ldots, |V|\}$ for each vertex $i \in V$. It is easily shown that for any given DAG it is possible to relabel the vertices so that parents always have a higher numbering that their respective children though such an ordering is not unique in general. For a given parent ordering we say that P obeys the parent ordered Markov property (PO) w.r.t. $\mathcal{G}$ if for every vertex i we have

$$i \perp \{i+1, \ldots, |V|\} \setminus pa(i)|pa(i).$$

It can be shown that if P has a density w.r.t. μ, then P obeys one of the directed Markov properties DF, DG, DL, PO if and only if it obeys all of them, i.e., the four Markov properties for DAGs are equivalent under mild conditions [13].

3 Gaussian directed acyclic graphical models

In this section we focus on multivariate Gaussian distributions which obey the directed Markov property w.r.t. a DAG $\mathcal{G}$. From now on and unless otherwise stated, we shall always assume without loss of generality that $\mathcal{G} = (V, E)$ is given in a parent ordering.

[5] see [13] for definition.

A Gaussian Bayesian network over $\mathcal{G}$ (or Gaussian DAG over $\mathcal{G}$), denoted by $\mathcal{N}(\mathcal{G})$, is the statistical model that consists of all multivariate Gaussian distributions $\mathrm{N}_m(\mu, \Sigma)$ which follow the directed Markov property w.r.t. $\mathcal{G}$ where $\mu \in \mathbb{R}^m$ and $\Sigma \in \mathrm{PD}_m(\mathbb{R})$, the set of $m \times m$ real positive definite matrices.

3.1 Linear recursive properties of Gaussian DAGs

Let $\mathbf{x} = (x_1, \ldots, x_m)^t$ be a random vector in $\mathbb{R}^m$ with the multivariate distribution $\mathrm{N}_m(0, \Sigma)$. Consider the system of linear recursive regression equations:

$$\begin{aligned} x_1 + \beta_{12}x_2 + \beta_{13}x_3 + \cdots + \beta_{1m}x_m &= \epsilon_1 \quad \text{or equivalently} \quad x_1 = -\beta_{12}x_2 - \beta_{13}x_3 - \cdots - \beta_{1m}x_m + \epsilon_1 \\ x_2 + \beta_{23}x_3 + \cdots + \beta_{2m}x_m &= \epsilon_2 \qquad\qquad\qquad\quad x_2 = -\beta_{23}x_3 - \cdots - \beta_{2m}x_m + \epsilon_2 \\ &\vdots \qquad\qquad\qquad\qquad\qquad\quad \vdots \\ x_m &= \epsilon_m \qquad\qquad\qquad\qquad\quad x_m = \epsilon_m, \end{aligned}$$

where $-\beta_{ij}$ is the partial regression coefficient of x_j ($j > i$) in the regression of x_i on its predecessors $x_{i+1}, \ldots, x_j, \ldots, x_m$. Now β_{ij} is zero if and only if $i \perp\!\!\!\perp \{i+1, \ldots, |V|\} \setminus pa(i) \mid pa(i)$. Hence the partial regression coefficient β_{ij} is zero if there does not exist an arrow from j to i, i.e., $j \notin pa(i)$, $j > i$. In addition, the residuals ϵ_i are normally distributed and mutually independent with mean zero and variance $\sigma^2_{i|pa(i)}$. We can rewrite the first system of equations in the form of a linear system $B\mathbf{x} = \boldsymbol{\epsilon}$, where B is the upper triangular matrix

$$B = \begin{pmatrix} 1 & \beta_{12} & \ldots & \beta_{1m} \\ 0 & 1 & \ldots & \beta_{2m} \\ 0 & \ldots & \ddots & \vdots \\ 0 & \ldots & 0 & 1 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}.$$

From this we obtain:

$$\begin{aligned} &\mathbb{V}ar[B\mathbf{x}] = \mathbb{V}ar[\epsilon] \\ \Rightarrow &B\Sigma B^t = diag(\sigma^2_{1|pa(1)}, \ldots, \sigma^2_{m-1|pa(m-1)}, \sigma^2_{mm}) =: D \\ \Rightarrow &\Sigma = B^{-1}D(B^t)^{-1} \\ \Rightarrow &\Sigma^{-1} = B^tD^{-1}B. \end{aligned} \tag{3.1}$$

Thus, if we define $L = B^t$, then $\Sigma^{-1} = LD^{-1}L^t$ is the so-called modified Cholesky decomposition of Σ^{-1}, in terms of the lower triangular matrix L and the diagonal matrix D^{-1}. Now consider a DAG denoted by $\mathcal{G} = (V, E)$. In [25] it has been shown that $\mathrm{N}_m(0, \Sigma)$ obeys the directed Markov property w.r.t. $\mathcal{G}$ if and only if $L_{ij} = 0$ whenever there is no arrow from i to j, i.e., $i \notin pa(j)$. Equation (3.1) above therefore gives a very convenient description of the Gaussian Bayesian network $\mathcal{N}(\mathcal{G})$. We explore this model in more detail below.

4 Parameterizations of $\mathscr{N}(\mathcal{G})$

In this subsection we discuss two parameterizations of the Gaussian Bayesian network $\mathscr{N}(\mathcal{G})$ that is of use in subsequent analysis. Let $a, b \subset V$ [6] and let $R^{a\times b}$ denote the real linear space of functions

$$T = ((i, j) \mapsto T_{ij}) : a \times b \to \mathbb{R}.$$

Each element of $\mathbb{R}^{a\times b}$ is called an $|a| \times |b|$ matrix. In particular, we define the space of symmetric matrices $S_a(\mathbb{R}) = \{T \in \mathbb{R}^{a\times a} : T_{ij} = T_{ji}\ \forall i, j \in a\}$, and the set of positive definite matrices as

$$\mathrm{PD}_a(\mathbb{R}) = \{T \in S_a(\mathbb{R}) :\ \xi^t T\xi > 0 : \forall \xi \in \mathbb{R}^a \setminus \{0\}\}.$$

Now let the sets a, b be a partition of the set V, then the positive definite matrix $\Sigma \in \mathrm{PD}_m(\mathbb{R})$ can be partitioned into block matrices as follows:

$$\begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix},$$

where $\Sigma_{aa} = (\Sigma_{ij})_{i,j\in a} \in \mathrm{PD}_a(\mathbb{R})$, $\Sigma_{bb} = (\Sigma_{jj})_{j,j\in b} \in \mathrm{PD}_b(\mathbb{R})$, $\Sigma_{ab} = (\Sigma_{ij})_{i\in a, j\in b} \in \mathbb{R}^{a\times b}$ and $\Sigma_{ba} = \Sigma_{ab}^t$. The Schur complement of the sub-matrix Σ_{aa} is defined as $\Sigma_{bb|a} = \Sigma_{bb} - \Sigma_{ba}(\Sigma_{aa})^{-1}\Sigma_{ab}$.

Remark 4.1. Throughout this paper, we usually suppress the notation for a principal sub-matrix Σ_{aa} and refer to it as Σ_a. We shall also use the convention Σ_a^{-1} for $(\Sigma_{aa})^{-1}$ and Σ^a for $(\Sigma^{-1})_{aa}$.

We now recall basic results from standard multivariate statistical theory. If $\mathbf{x} \sim \mathrm{N}_m(\mu, \Sigma)$, $\mu \in \mathbb{R}^V$, then $\mathbf{x}_a \sim \mathrm{N}_a(\mu_a, \Sigma_a)$ and for any $x_b \in \mathbb{R}^b$, the conditional distribution of $\mathbf{x}_a|\mathbf{x}_b = x_b$ is given by $\mathrm{N}_a(\mu_a + \Sigma_{ab}\Sigma_b^{-1}(x_b - \mu_b), \Sigma_{aa|b})$, i.e., $\Sigma_{ab}\Sigma_b^{-1}$ is the regression coefficient of $\mathbf{x}_b$ in the regression of $\mathbf{x}_a$ on $\mathbf{x}_b$, and $\Sigma_{aa|b}$ is the conditional variance of the residual. More generally, for a partition a, b, c of V denote the corresponding block partition matrices of Σ and Σ^{-1} as follows:

$$\Sigma = \begin{pmatrix} \Sigma_a & \Sigma_{ab} & \Sigma_{ac} \\ \Sigma_{ba} & \Sigma_b & \Sigma_{bc} \\ \Sigma_{ca} & \Sigma_{cb} & \Sigma_c \end{pmatrix} \text{ and } \Sigma^{-1} = \begin{pmatrix} \Sigma^a & \Sigma^{ab} & \Sigma^{ac} \\ \Sigma^{ba} & \Sigma^b & \Sigma^{bc} \\ \Sigma^{ca} & \Sigma^{cb} & \Sigma^c \end{pmatrix}.$$

Then the partial covariance matrix of both of $\mathbf{x}_a$ and $\mathbf{x}_c$ given $\mathbf{x}_b = x_b$ is denoted as $\begin{pmatrix} \Sigma_{aa|b} & \Sigma_{ac|b} \\ \Sigma_{ca|b} & \Sigma_{cc|b} \end{pmatrix}$, where $\Sigma_{ac|b} = \Sigma_{ac} - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{bc}$, and $\Sigma_{ca|b} = \Sigma'_{ac|b}$. In particular,

$$a \perp\!\!\!\perp c|b \text{ implies that } \Sigma_{ac|b} = 0 \Leftrightarrow \Sigma_{ac} = \Sigma_{ab}\Sigma_b^{-1}\Sigma_{bc}. \tag{4.1}$$

Next, let $\mathrm{PD}_{\mathcal{G}}$ denote the set of positive definite matrices Σ in $\mathrm{PD}_m(\mathbb{R})$ such that $\mathrm{N}_m(\mu, \Sigma) \in \mathscr{N}(\mathcal{G})$ for every $\mu \in \mathbb{R}^m$. Clearly, $\mathrm{N}_m(\mu, \Sigma) \in \mathscr{N}(\mathcal{G})$ if and only if $\mathrm{N}_m(0, \Sigma) \in \mathscr{N}(\mathcal{G})$. Therefore, without loss of generality, we shall only consider centered Gaussian distributions $\{\mathrm{N}_m(0, \Sigma) :\ \Sigma \in \mathrm{PD}_{\mathcal{G}}\} \subset \mathscr{N}(\mathcal{G})$. For convenience however, we shall still denote the submodel $\{\mathrm{N}_m(0, \Sigma) :\ \Sigma \in \mathrm{PD}_{\mathcal{G}}\}$ by $\mathscr{N}(\mathcal{G})$.

[6] Note that under-case alphabets are used to denote subsets of V.

4.1 D-parametrization

A parameterization ("D-parameterization") of $\mathcal{N}(\mathcal{G})$ can be obtained by using the recursive factorization property of the Gaussian densities in $\mathcal{N}(\mathcal{G})$. First we recall the following notation from [1].

Notation. For each $i \in V$ let

$$\begin{array}{lll} \prec i \succ = pa(i), & [i \succ = \{i\} \times pa(i), & \prec i] = pa(i) \times \{i\}, \\ \not\prec i \not\succ = \{j : j > i\} \setminus pa(i), & [i \not\succ = \{i\} \times \not\prec i \not\succ, & \prec i \not\succ = \prec i \succ \times \not\prec i \not\succ \\ \preceq i \succeq = fa(i) & & \end{array}$$

By applying the directed factorization property (DF) of $\mathrm{N}_m(0, \Sigma) \in \mathcal{N}(\mathcal{G})$ we have

$$d\mathrm{N}_m(0,\Sigma)(x) = \prod_{i \in V} d\mathrm{N}(\mu_{i|pa(i)}, \Sigma_{i|pa(i)})(x_i|x_{pa_i}) = \prod_{i \in V} d\mathrm{N}(\Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ} x_{\prec i\succ}, \Sigma_{ii|\prec i\succ})(x_i), \qquad (4.2)$$

for each $x = (x_i)_{i \in V} \in \mathbb{R}^m$. Note that $\mathrm{N}(\Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ} x_{\prec i\succ}, \Sigma_{ii|\prec i\succ})(x_i)$ is the conditional distribution $\mathbf{x}_i|\mathbf{x}_{\prec i\succ} = x_{\prec i\succ}$. Furthermore, using the exact functional form for the densities of the Gaussian distributions in Equation (4.2), we obtain the following expression

$$\mathrm{tr}(\Sigma^{-1} x x^t) = \sum_{i \in V} \mathrm{tr}\left(\Sigma^{-1}_{ii|\prec i\succ}(x_i - \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ} x_{\prec i\succ})(x_i - \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ} x_{\prec i\succ})^t\right). \qquad (4.3)$$

It is shown in [1] that $\Sigma \in \mathrm{PD}_{\mathcal{G}}$ if and only if $\Sigma \in \mathrm{PD}_m(\mathbb{R})$ and satisfies Equation (4.3) for all $x \in \mathbb{R}^m$. On the other hand, by the parent ordered Markov property (PO) of $\mathrm{N}_m(0, \Sigma)$ we have $\Sigma \in \mathrm{PD}_{\mathcal{G}}$ if $i \perp\!\!\!\perp \not\prec i \not\succ \mid \prec i \succ$ (or equivalently $i \perp \{i+1, \ldots, m\} \setminus pa(i) | pa(i)$). Another characterization given by [1] for $\Sigma \in \mathrm{PD}_{\mathcal{G}}$ is that $\Sigma \in \mathrm{PD}_m(\mathbb{R})$ and

$$\Sigma_{[i\not\succ} = \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ}\Sigma_{\prec i\not\succ}, \ \forall i \in V. \qquad (4.4)$$

Using the insights above and defining $\Xi_{\mathcal{G}} = \times_{i \in V}\left(\mathbb{R}_+ \times \mathbb{R}^{\prec i]}\right)$, it can be shown that the mapping

$$\left(\Sigma \mapsto \times_{i \in V}(\Sigma_{ii|\prec i\succ}, \Sigma^{-1}_{\prec i\succ}\Sigma_{\prec i]})\right) : \mathrm{PD}_{\mathcal{G}} \to \Xi_{\mathcal{G}}$$

is a bijection. In order to construct the inverse of this mapping let $\times_{i \in V}(\lambda_i, \beta_{[i\succ})$ be a typical element in $\Xi_{\mathcal{G}}$, with the convention that $\beta_{[i\succ} = 0$ whenever $\prec i \succ = pa(i) = \emptyset$. The corresponding Σ can be recursively constructed starting from the largest index m, by setting

$$\begin{array}{ll} i) & \Sigma_{ii} = \lambda_i + \beta^t_{\prec i]}\Sigma_{\prec i\succ}\beta_{\prec i]}, \\ ii) & \Sigma_{\prec i]} = \Sigma_{\prec i\succ}\beta_{\prec i]}, \text{ and} \\ iii) & \Sigma_{[i\not\succ} = \Sigma_{[i\succ}\Sigma^{-1}_{\prec i\succ}\Sigma_{\prec i\not\succ}, \text{ according to Equation (4.4)} \end{array} \qquad (4.5)$$

The reader is referred to [1] for greater detail, where in addition, it is shown that the above inverse mapping yields a positive definite matrix in $\mathrm{PD}_m(\mathbb{R})$, and consequently in $\mathrm{PD}_{\mathcal{G}}$. We shall revisit the "D-parameterization" mapping in subsequent sections.

4.2 Cholesky parameterization

An alternative parameterization ("Cholesky-parameterization") of $\mathcal{N}(\mathcal{G})$ can be obtained by using the Cholesy decomposition of Σ^{-1}. From §3.1 it is clear that $\mathcal{N}(\mathcal{G})$ is isomorphic to the following family of distributions:

$$\mathcal{N}(\mathcal{G}) \cong \{\mathrm{N}_m(0, (L^t)^{-1}DL^{-1}) : \ (D, L) \in \Theta_{\mathcal{G}}\},$$

where $\Theta_{\mathcal{G}} = \mathcal{D}_m^+ \times \mathcal{L}_{\mathcal{G}}$, with $\mathcal{L}_{\mathcal{G}}$ being the set of lower triangular matrices L in $\mathbb{R}^{m\times m}$ such that

$$L_{ij} = \begin{cases} 1 & i = j \\ L_{ij} & i > j \text{ and } i \in pa(j) \\ 0 & \text{otherwise} \end{cases}$$

and $\mathcal{D}_m^+$ the set of diagonal matrices in $\mathbb{R}^{m\times m}$ such that $D_{ii} > 0$. Note that $\Sigma^{-1} = LD^{-1}L^t$ and that if $i \notin pa(j)$ then $L_{ij} = 0$.

Remark 4.2. Note that the D-parameterization and and Cholesky parametrization of a Gaussian DAG model $\mathcal{N}_{\mathcal{G}}$ are essentially the same, as they both encode the partial regression coefficients and residuals in the system of regression equations described in Subsection 3.1. The difference between the two parameterizations are subtle. The $\beta_{[i>}$ elements of the D-parameterization constitute the non-zero off-diagonal elements of the columns of the L matrix in the Cholesky parameterization whereas the λ_i in the D-parameterization is the D_{ii} in the Cholesky parameterization. The latter parameterization is essentially a matrix version of the former. The zero and non-zero elements of the L matrix in the Cholesky parameterization encodes the graph corresponding to the DAG, whereas in the D-parameterization the information regarding the parents of each vertex is assumed to implicitly accompany the parameters. We shall see later that each parametrization has its respective advantage. While computations involving the D-parameterization are often easier, the Cholesky decomposition intuitively encodes the structure of the underlying DAG and provides a natural description of the Wishart DAG densities introduced in the paper.

4.3 Covariance manifolds

In this subsection we introduce parameter spaces that correspond to general Gaussian conditional independence models. We note that the spaces defined here are for more general conditional independence models and not just for DAG models. The reason for this is three fold. The first stems from a desire to give a unified treatment of the parameter spaces for DAGs and undirected graphical models (UGs). Second, a more general definition will allow us to compare the distributions introduced in this paper for DAGs to those that have been previously proposed in the literature for UGs. Third, for special class of graphs (such as decomposable or perfect graphs), DAGs and UGs coincide in terms of the conditional independences that such graphs can encode. In these instances, the ability to exploit the equivalence between DAGs and UGs requires definitions of parameter spaces that are more general.

We shall use the term general Markov model (or GM model), to refer to the statistical models which are defined in terms of a set of conditional independence constraints. Two GM models are said to be "Markov equivalent" if every distribution that satisfies the required conditional independence assumptions in one model will satisfy those in the other and vice versa. An important example of Markov equivalence classes are the class of perfect DAGs and the class of decomposable undirected graphs, i.e., for a perfect DAG $\mathcal{G}$ the class of GM models over $\mathcal{G}$ coincides with that of GM models over the decomposable graph $\mathcal{G}^{\mathrm{u}}$, the undirected version of $\mathcal{G}$ (see [13] for details).

One important subclass of GM models over an arbitrary graph $\mathcal{G} = (V, E)$, directed or undirected is the general Gaussian graphical model over $\mathcal{G}$, denoted by $\mathcal{N}(\mathcal{G})$ and represented by the class of multivariate normal distributions $\mathrm{N}_m(0, \Sigma)$ obeying the corresponding Markov properties w.r.t. $\mathcal{G}$. More formally, let $\mathcal{G} = (V, E)$ be a graph (directed or undirected) and $\mathcal{N}(\mathcal{G})$ the general Gaussian graphical model over $\mathcal{G}$. Consider the following definitions.

Definition 4.1. (*a*) $\mathrm{PD}_{\mathcal{G}}$ is the set of positive definite matrices Σ in $\mathrm{PD}_m(\mathbb{R})$ such that $\mathrm{N}_m(0, \Sigma) \in \mathcal{N}(\mathcal{G})$.

(*b*) $\mathrm{P}_{\mathcal{G}}$ is the set of positive definite matrices Ω such that $\Omega^{-1} \in \mathrm{PD}_{\mathcal{G}}$.

(*c*) $\mathrm{Z}_{\mathcal{G}}$ is the real linear space of $m \times m$ symmetric matrices T of dimension m such that $T_{ij} = T_{ji} = 0$ if (i, j) is not in E.

(*d*) $\mathrm{I}_{\mathcal{G}}$ is the real linear space of symmetric functions $\Gamma = (\Gamma : (i, j) \mapsto \Gamma_{ij}) : E^{\mathrm{u}} \to \mathbb{R}$, i.e., $\Gamma_{ij} = \Gamma_{ji}$. An element $\Gamma \in \mathrm{I}_{\mathcal{G}}$ is called a $\mathcal{G}$-incomplete (symmetric) matrix.

(*e*) $\mathrm{Q}_{\mathcal{G}}$ is the set of $\mathcal{G}$-incomplete matrices $\Gamma \in \mathrm{I}_{\mathcal{G}}$ such that Γ_c is positive definite for each clique $c \in \mathscr{C}_{\mathcal{G}}$. Elements of $\mathrm{Q}_{\mathcal{G}}$ are said to be *partially positive definite* matrices over $\mathcal{G}$.

Remark 4.3. Note that $\mathrm{I}_{\mathcal{G}}$ is naturally a real linear subspace of $\mathrm{S}_m(\mathbb{R})$. If $\mathcal{G} = (V, E)$ is a *DAG*, then $\mathrm{I}_{\mathcal{G}}$ is an $|E|-$dimensional linear subspace of $\mathrm{S}_m(\mathbb{R})$.

Remark 4.4. We emphasize that these definitions are more general than those introduced in [15] and used in [19], as $\mathcal{G}$ can now be either directed or undirected. However, when the graph $\mathcal{G}$ is undirected the two definitions do coincide, and thus the definitions above are in some sense extended versions of those in [15]. In particular, for a undirected decomposable $\mathcal{G}$ we have $\mathrm{P}_{\mathcal{G}} = \{\Omega \in \mathrm{PD}_m(\mathbb{R}) : \omega_{ij} = 0 \text{ if } (i, j) \notin E\}$. Moreover, when $\mathcal{G}$ is a perfect DAG then $\mathrm{P}_{\mathcal{G}}$ ($\mathrm{Q}_{\mathcal{G}}$) and $\mathrm{P}_{\mathcal{G}^{\mathrm{u}}}$ ($\mathrm{Q}_{\mathcal{G}^{\mathrm{u}}}$) are identical due to the Markov equivalence property of perfect DAGs and decomposable undirected graphs.

We also introduce additional notation that is required in subsequent sections:

Notation. Let $\mathcal{G} = (V, E)$ be a DAG. For a symmetric matrix $T \in \mathrm{S}_m(\mathbb{R})$, we denote by T^E the projection of T on $\mathrm{I}_{\mathcal{G}}$. The projection mapping $(T \mapsto T^E) : \mathrm{S}_m(\mathbb{R}) \to \mathrm{I}_{\mathcal{G}}$ is denoted by $\mathrm{proj}_{\mathcal{G}}$.

5 Generalized Wishart laws for DAG Models

In this section we introduce new classes of matrix variate distributions for parameters of interests in the Gaussian DAG model.

5.1 The DAG Wishart distribution on $\Theta_{\mathcal{G}}$

The modified Cholesky decomposition provides a natural description of Gaussian DAG models and hence we start by developing distributions on the space $\Theta_{\mathcal{G}}$ (as defined in §4). It is assumed henceforth, unless otherwise stated, that $\mathcal{G} = (V, E)$ is a DAG and that the vertices in $V = \{1, \ldots, m\}$ are parent-ordered[7] i.e., $pa(i) \subset \{i+1, \ldots, m\}$ for each $i = 1, \ldots, m-1$. Recall that the Gaussian DAG model associated with $\mathcal{G}$ is the family of distributions

$$\mathcal{N}(\mathcal{G}) = \{\mathrm{N}_m(\mathbf{0}, \Sigma) : \Sigma \in \mathrm{PD}_{\mathcal{G}}\} \cong \{\mathrm{N}_m(\mathbf{0}, (L^{-1})^T D L^{-1}) : (D, L) \in \Theta_{\mathcal{G}}\}.$$

Consider the family of measures on $\Theta_{\mathcal{G}}$ with density $\left(\text{w.r.t. } \prod_{i>j,(i,j)\in E} dL_{ij} \prod_{i=1}^m dD_{ii}\right)$

$$\tilde{\pi}_{U,\alpha}(D, L) = \exp\{-\frac{1}{2}\mathrm{tr}((LD^{-1}L^t)U)\} \prod_{i=1}^m D_{ii}^{-\frac{1}{2}\alpha_i}, \quad (D, L) \in \Theta_{\mathcal{G}}. \tag{5.1}$$

This family of measures is parameterized by a positive definite matrix U and a vector $\alpha \in \mathbb{R}^m$ with non-negative entries. Let

$$z_{\mathcal{G}}(U, \alpha) := \int_{\Theta_{\mathcal{G}}} \tilde{\pi}_{U,\alpha}\, dLdD = \int_{\Theta_{\mathcal{G}}} \exp\{-\frac{1}{2}\mathrm{tr}(LD^{-1}L^tU\} \prod_{i=1}^m D_{ii}^{-\frac{1}{2}\alpha_i} dLdD.$$

If $z_{\mathcal{G}}(U, \alpha) < \infty$, then $\tilde{\pi}_{U,\alpha}$ can be normalized to obtain a probability measure. A necessary and sufficient condition for the existence of a normalizing constant for ANY arbitrary DAG is obtained in the following theorem.

Theorem 5.1. *Let $dL := \prod_{(i,j)\in E, i>j} dL_{ij}$ and $dD := \prod_{i=1}^m dD_{ii}$ denote, respectively, the canonical Lebesgue measures on $\mathcal{L}_{\mathcal{G}}$ and $\mathbb{R}_+^m$ and let $pa_i := |pa(i)|$. Then,*

$$\int_{\Theta_{\mathcal{G}}} \exp\{-\frac{1}{2}\mathrm{tr}(LD^{-1}L^tU)\} \prod_{i=1}^m D_{ii}^{-\frac{1}{2}\alpha_i} dLdD < \infty$$

if and only if

$$\alpha_i > pa_i + 2 \quad \forall i = 1, \ldots, m.$$

Furthermore, in this case

$$z_{\mathcal{G}}(U, \alpha) = \prod_{i=1}^m \frac{\Gamma\left(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1\right) 2^{\frac{\alpha_i}{2}-1} (\sqrt{\pi})^{pa_i} \det(U_{<i>})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-\frac{3}{2}}}{\det(U_{\leq i\geq})^{\frac{\alpha_i}{2}-\frac{pa_i}{2}-1}}. \tag{5.2}$$

[7] We emphasize here that unlike in the decomposable concentration and covariance graph setting (where the existence of an ordering is important either for the perfect order of cliques and separators, or to preserve zeros), existence of such an ordering is not necessary in the DAG setting, since a parent ordering is always available for a DAG.

Proof. The proof of this theorem is given in the Appendix/Supplemental section. □

Note that the expression above for the normalizing constant bears close resemblance to Corollary 3 of [9] due to the Markov equivalence of DAGs and covariance graph models when is $\mathcal{G}$ is a homogeneous graph. A through investigation of this parallel requires more tools and is undertaken in the sequel to this paper - see [3]. At a first glance the distribution defined in Equation (5.1) appears to be the same as the covariance Wishart distributions defined in [9], but a closer in-depth look reveals that they are different. Note that the product $LD^{-1}L^t$ features in Equation (5.1) whereas the expression $(LDL^t)^{-1}$ features in the density of the covariance Wishart distributions defined in [9]. More importantly however, the closed form result above is valid for ALL DAGs and not restricted to any specific subclass of graphs such as perfect(or decomposable graphs) as in the treatment of undirected graphs [15, 19], or, the class of homogeneous graphs as in the treatment of covariance graphs [9]. As will be seen later this property has significant consequences for Bayesian model selection in high dimensional settings.

Definition 5.1. The normalized version of $\widetilde{\pi}_{U,\alpha}$, denote by $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$, will be referred to as the "DAG Wishart" distribution on $\Theta_{\mathcal{G}}$ with shape parameter $U \in \mathrm{PD}_m(\mathbb{R})$ and multivariate scale parameter $\alpha = (\alpha_i, \ldots, \alpha_m)^t \in \mathbb{R}^m$. The density of $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$, when $\alpha_i > pa_i + 2$ for every $i \in V$, is given as follows:

$$\pi^{\Theta_{\mathcal{G}}}_{U,\alpha} = z_{\mathcal{G}}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(LD^{-1}L^tU)\} \prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i}, \quad (D,L) \in \Theta_{\mathcal{G}},$$

where the normalizing constant $z_{\mathcal{G}}(U,\alpha)$ is given in Equation (5.2).

We now proceed to demonstrate that the class of matrix-variate DAG Wishart probability distributions defined above has important statistical uses. The following lemma shows that the family $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$ is standard conjugate for Gaussian DAG models.

Lemma 5.1. *Let $\mathcal{G}$ be an aribtrary DAG and let $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ be an i.i.d. sample from $\mathrm{N}_m(\mathbf{0}, (L^{-1})^t D L^{-1})$, where $(D,L) \in \Theta_{\mathcal{G}}$. Let $S = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Y}_i\mathbf{Y}_i^t$ denote the empirical covariance matrix. If the prior distribution on (D,L) is $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$, then the posterior distribution of (D,L) is given by $\pi^{\Theta_{\mathcal{G}}}_{\widetilde{U},\widetilde{\alpha}}$, where $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (n+\alpha_1, n+\alpha_2, \cdots, n+\alpha_m)$.*

Proof. The proof is given in the Appendix/Supplemental section. □

Remark 5.1. The case when the observations do not have mean zero (i.e., when $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ are i.i.d. $\mathrm{N}_m(\mu, \Sigma)$, with $\mu \in \mathbb{R}^m$, $\Sigma \in \mathrm{PD}_{\mathcal{G}}$) can be handled in a similar manner by noting that the sample covariance matrix S is a sufficient statistic for Σ and the fact that $nS \sim \mathrm{W}_m(n-1, \Sigma)$.

6 Hyper Markov properties

This section explores the distributional properties of the $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$ family in detail. In particular, we demonstrate that there is deep and useful structure in the DAG Wishart distributions introduced in Section 5, with important implications for statistical inference.

6.1 Strong directed hyper Markov properties of $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$

The conceptual foundations of hyper Markov properties were laid in [7] and the reader is referred to [15] for a brief overview. Consider the Gaussian DAG model $\mathcal{N}(\mathcal{G})$ with the parameter space $\Theta_{\mathcal{G}}$. The elements of $\mathcal{N}(\mathcal{G})$ are of the form $\mathrm{N}_m(0, (L^{-1})^T D L^{-1})$, such that $(D, L) \in \Theta_{\mathcal{G}}$. Recall that if $\mathbf{x} \sim \mathrm{N}_m(0, \Sigma)$, then for each $i \in V$ the distribution of $\mathbf{x}_{ii|<i>}$ is parametrized by $(\Sigma_{ii|<i>}, \Sigma^{-1}_{<i>}\Sigma_{<i]})$. Note furthermore that these parameters are related to the Cholesky parameterization as follows: $D_{ii} = \Sigma_{ii|<i>}$ and $L_{<i]} = -\Sigma^{-1}_{<i>}\Sigma_{<i]}$ (see [1, Proposition 11.1]). The following theorem establishes the strong directed hyper Markov property for the $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$ family of DAG Wishart distributions.

Theorem 6.1. *Let $\mathcal{G}$ be an arbitrary DAG and $(D, L) \sim \pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$. Then $\{(D_{ii}, L_{<i]}) : i = 1, \ldots, m\}$ are mutually independent. Moreover,*

$$D_{ii} \sim IG(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1, \frac{1}{2}U_{ii|<i>}), \text{ and} \tag{6.1}$$

$$L_{<i]}|D_{ii} \sim \mathrm{N}_{pa_i}(-U^{-1}_{<i>}U_{<i]}, D_{ii}U^{-1}_{<i>}). \tag{6.2}$$

Proof. The proof is given in the Appendix/Supplemental section. □

Theorem 6.1 yields the marginal density of D, whereas for the L parameter, only the conditional distribution given the D are given. Hence we now proceed to derive the marginal density of L.

Corollary 6.1. *Let $\mathcal{G}$ be an arbitrary DAG and suppose $(L, D) \sim \pi^{\Theta}_{U,\alpha}$. Then the density of L w.r.t. $dL = \prod_{i=1}^m dL_{<i]}$ is given by*

$$\prod_{i=1}^{m} c_i \left[1/2 U_{ii|<i>} + (L_{<i]} + U^{-1}_{<i>}U_{<i]})^t U_{<i>}(L_{<i]} + U^{-1}_{<i>}U_{<i]})\right]^{-\alpha_i/2+1},$$

where each c_i is given by

$$\frac{\det(U_{<i>})^{1/2}(U_{ii|<i>})^{\alpha_i/2-pa_i/2-1}\Gamma(\alpha_i/2-1)}{2^{\alpha_i/2-1}\pi^{pa_i/2}\Gamma(\alpha_i/2-pa_i/2-1)}. \tag{6.3}$$

In particular, for each i, $L_{<i]}$ has a multivariate t-distribution:

$$L_{<i]} \sim \mathbf{t}_{pa_i}\left(-U^{-1}_{<i>}U_{<i]}, (\alpha_i/2 - pa_i/2 - 1)U_{ii|<i>}U^{-1}_{<i>}, \alpha_i - pa_i - 2\right),$$

i.e., $L_{<i]}$ has a pa_i-variate t-distribution with mean parameter $-U^{-1}_{<i>}U_{<i]}$, scale parameter $(\alpha_i/2 - pa_i/2 - 1)U_{ii|<i>}U^{-1}_{<i>}$ and degrees of freedom $\nu_i := \alpha_i - pa_i - 2$.

Proof. The proof is given in the Appendix/Supplemental section. □

6.2 Alternative method for deriving the DAG Wishart distribution

In light of Theorem 6.1 and following the general approach in [14] we show below that one can arrive at the DAG Wishart distributions $\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}$ in an alternative way. Consider a general setting and suppose that $\{P_\theta :\in \Theta\}$ is a family of directed Markov fields over a DAG $\mathcal{G} = (V, E)$. For each vertex $i \in V$ let θ_i (or more accurately $\theta_{i|pa(i)}$) be the corresponding parameter of the conditional probability of $\mathbf{x}_i|\mathbf{x}_{pa(i)}$ and let $\Theta_i = \{\theta_i : \theta \in \Theta\}$. Therefore, under a re-parameterization, we may assume that $\Theta = \times_{i=1}^m \Theta_i$. Now suppose that for each vertex i, a prior $\pi_i(\theta_i)$ is specified. Now under the global independence assumption, i.e., the assumption that the parameters $\{\theta_i : i \in V\}$ are apriori independent random variables, we have $\pi(\theta) = \prod_{i=1}^m \pi_i(\theta_i)$. Note that global independence assumption is the same as assuming the strong hyper Markov property on the parameter θ (see [7]).

Now consider a Gaussian DAG model given as $\{\mathrm{N}_m(0, \Sigma) : \Sigma \in \mathrm{PD}_{\mathcal{G}})\}$. Recall that if $\mathbf{x} \sim \mathrm{N}_m(0, \Sigma)$, then the distribution of $\mathbf{x}_i|\mathbf{x}_{<i>}$ is parameterized by $(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]})$. This suggests a re-parametrization of the original family of distributions using the D-parameterization introduced in Section 4.1

$$\Xi_{\mathcal{G}} = \{\times_{i=1}^m (\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]}) : \Sigma \in \mathrm{PD}_{\mathcal{G}}\}.$$

Since $\mathbf{x}_i|\mathbf{x}_{<i>} \sim \mathrm{N}(\Sigma_{[i>}\Sigma_{<i>}^{-1}x_{<i>}, \Sigma_{ii|<i>})$, a natural approach to prior specification for the parameter set $(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]})$ is the standard conjugate prior, i.e., the inverse-gamma distribution for $\Sigma_{ii|<i>}$ and a Gaussian distribution for $\Sigma_{<i>}^{-1}\Sigma_{<i]}|\Sigma_{ii|<i>}$. More precisely if

$$\Sigma_{ii|<i>} \sim IG(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1, \frac{1}{2}U_{ii|<i>}), \text{ and}$$

$$\Sigma_{<i>}^{-1}\Sigma_{<i]}|\Sigma_{ii|<i>} \sim \mathrm{N}_{pa_i}(U_{<i>}^{-1}U_{<i]}, \Sigma_{ii|<i>}U_{<i>}^{-1}),$$

then the global independence assumption implies that the distribution of $\times_{i=1}^m (\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]})$ is proportional to

$$\exp\{-\frac{1}{2}\sum_{i\in V}\Sigma_{ii|<i>}^{-1}(\Sigma_{<i>}^{-1}\Sigma_{<i]} - U_{<i>}^{-1}U_{<i]})^t U_{<i>}(\Sigma_{<i>}^{-1}\Sigma_{<i]} - U_{<i>}^{-1}U_{<i]}) + \Sigma_{ii|<i>}^{-1}U_{ii|<i>}\}\prod_{i\in V}\Sigma_{ii|<i>}^{-\frac{1}{2}\alpha_i}.$$

It is now clear from the proof of Theorem 6.1 that the DAG Wishart distributions $\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}$ defined in Equation (12.2) is nothing but the image of this distribution under the map $\phi^{-1} : \Xi_{\mathcal{G}} \to \Theta_{\mathcal{G}}$.

Remark 6.1. We note that though the class of DAG Wishart distributions can be derived in an equivalent way using the general approach in [7], the latter approach does not however immediately give a means to specify hyper-parameters that will exactly correspond to our DAG Wishart distributions. Hence it is important to note that the equivalence is easier to recognize once hyper Markov properties for our DAG Wishart distributions have been established.

7 Laplace transform, expected value and posterior mode

In this section we compute Laplace transforms, expected values and posterior models for the class of DAG Wishart distribution defined in this paper.

7.1 Laplace transforms

We start with computing the Laplace transform of $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$ by exploiting the results established in Theorem 6.1. First a preliminary result on the Laplace transform of a Gaussian inverse Gamma distribution is required.

Lemma 7.1. *Suppose* $(\lambda, \mathbf{x})$ *is a random variable with Gaussian-inverse gamma distribution:*

$$\begin{aligned} \mathbf{x}|\lambda &\sim \mathrm{N}_p(\mu, \lambda\Psi), \quad \mu \in \mathbb{R}^p, \Psi \in \mathrm{PD}_p(\mathbb{R}); \\ \lambda &\sim \mathrm{IG}(\nu, \eta). \end{aligned}$$

Then the Laplace transform of $(\lambda, \mathbf{x})$ *at* $(\xi, u) \in \mathbb{R}_+ \times \mathbb{R}^p_+$ *is*

$$\frac{2}{\Gamma(\nu)} \exp\{u^t\mu\} \left(\eta(\xi - \frac{1}{2}u^t\Psi u)\right)^{\frac{1}{2}\nu} K_\nu\left(2\sqrt{\eta(\xi - \frac{1}{2}u^t\Psi u)}\right),$$

where $K_\nu(\cdot)$ *is the modified Bessel function of the second type and* $\xi - \frac{1}{2}u^t\Psi u$ *is assumed to be positive.*

Proof. The proof is given in the Appendix/Supplemental section. □

Proposition 7.1. *The Laplace transform of* $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$ *at a typical point* $\times_{i=1}^m(\xi_i, z_{<i]}) \in \Xi_{\mathcal{G}}$ *is given by*

$$\mathcal{L}_{\Xi_{\mathcal{G}}}(\times_{i=1}^m(\xi_i, z_{<i]})) := 2^m \prod_{i=1}^m \left(\frac{1}{\Gamma(r_i)} \exp\{z^t_{<i]}\mu_{<i]}\} \left(\eta_i(\xi_i - \frac{1}{2}z^t_{<i]}\Psi_{<i>}z_{<i]})\right)^{\frac{1}{2}r_i} K_{r_i}\left(2\sqrt{\eta_i(\xi_i - \frac{1}{2}z^t_{<i]}\Psi_{<i>}z_{<i]})}\right)\right),$$

where $r_i = \frac{\alpha_i}{2} - \frac{pa_i}{2} - 1$, $\eta_i = \frac{1}{2}U_{ii|<i>}$, $\mu_{<i]} = -U^{-1}_{<i>}U_{<i]}$, $\Psi_{<i>} = U^{-1}_{<i>}$, *and* $\xi_i - \frac{1}{2}z^t_{<i]}\Psi_{<i>}z_{<i]}$ *are assumed to be positive for each* i.

Proof. Let $\times_{i=1}^m(\lambda_i, \beta_{<i]}) \sim \pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$. Theorem 6.1 implies that the finite sequence of random variables $(\lambda_i, \beta_{<i]})$ are independent and each has a Gaussian-inverse gamma distribution as given by Equation (6.1) and Equation (6.2). It therefore suffices to compute the Laplace transform of each random vector $(\lambda_i, \beta_{<i]})$ individually. The Laplace transform of $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$ now follows immediately from Lemma 7.1. □

We now proceed to give the Laplace transform of $\pi^{\Theta}_{U,\alpha}$.

Corollary 7.1. *The Laplace transform of* $\pi^{\Theta}_{U,\alpha}$ *at* $(\Lambda, Z) \in \Theta_{\mathcal{G}}$ *is given by*

$$\left(\frac{2}{e}\right)^m \prod_{i=1}^m \left(\frac{1}{\Gamma(r_i)} \exp\{z^t_{<i]}\mu_{<i]}\} \left(\eta_i(\xi_i - \frac{1}{2}z^t_{<i]}\Psi_{<i>}z_{<i]})\right)^{\frac{1}{2}r_i} K_{r_i}\left(2\sqrt{\eta_i(\xi_i - \frac{1}{2}z^t_{<i]}\Psi_{<i>}z_{<i]})}\right)\right)$$

Proof. The proof is given in the Appendix/Supplemental section. □

7.2 Expected values

We now proceed to compute the expected values of our priors. First some necessary notation is introduced: Suppose $a, b \subset V$ and $A \in \mathbb{R}^{a\times b}$ a matrix of size $|a| \times |b|$. Then define $(A)^0 \in \mathbb{R}^{V\times V}$ by

$$(A)^0_{ij} = \begin{cases} A_{ij} & i \in a, j \in b \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, if $L_{<i]}$ is a vector in $\mathbb{R}^{<i]}$, then we consider

$$\begin{pmatrix} 1 \\ L_{<i]} \end{pmatrix}$$

as a vector in $\mathbb{R}^{\preceq i]}$ with 1 in ii position.

Now recall from 6.1 that $L_{<i]}$ has a multivariate t-distribution. This result readily allows us to compute the mean and covariance of the random elements of L. They are given as follows:

$$\mathbb{E}(L_{<i]}) = -U^{-1}_{<i>} U_{<i]} \text{ and } \mathbb{V}ar(L_{<i]}) = \frac{\nu_i^2}{2\nu_i - 4} U_{ii|<i>} U^{-1}_{<i>}.$$

Consequently, if $A := \{1, i_2, \ldots, i_r\} \subset V$ is the set of vertices i such that $pa(i) \neq \emptyset$, then $\mathbb{E}(\times_{i\in A} L_{<i]}) = -\times_{i\in A} U^{-1}_{<i>} U_{<i]}$. This can be expressed in matrix form as follows:

$$\mathbb{E}[L] = \mathbb{E}[\sum_{j=1}^{m} (L_{\cdot j})^0] = \sum_{j=1}^{m} \begin{pmatrix} 1 \\ \mathbb{E}[L_{<i]}] \end{pmatrix}^0 = \sum_{j=1}^{m} \begin{pmatrix} 1 \\ -U^{-1}_{<i>} U_{<i]} \end{pmatrix}^0.$$

The expression for $\mathbb{V}ar(\times_{i\in A} L_{<i]})$ is given by the block diagonal matrix

$$\begin{pmatrix} \frac{\nu_1^2}{2\nu_1-4} U_{11|<1>} U^{-1}_{<1>} & 0 & \cdots & 0 \\ 0 & \frac{\nu_{i_2}^2}{2\nu_{i_2}-4} U_{i_2 i_2|<i_2>} U^{-1}_{<i_2>} & & \\ \vdots & & \ddots & \\ 0 & & & \frac{\nu_{i_r}^2}{2\nu_{i_r}-4} U_{i_r i_r|<i_r>} U^{-1}_{<i_r>} \end{pmatrix}.$$

The expected value of D can also be easily computed using the result in Equation (6.1). Under the Cholesky decomposition parameterization we have $\mathbb{E}[D] = \text{Diag}\left(\frac{U_{ii|<i>}}{\alpha_i - pa_i - 4} : i \in V\right)$.

7.3 Posterior modes

We now proceed to compute the posterior mode of $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$ as this is often a useful quantity in Bayesian inference. The computation of the posterior modes under other parameterizations

follow from similar calculations. First let us compute the mode of $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$. Recall that from Equation (12.4) the density of $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$ is proportional to

$$\exp\{-\frac{1}{2}\sum_{i\in V}\lambda_i^{-1}(\beta_{<i]}+U^{-1}_{<i>}U_{<i]})^t U_{<i>}(\beta_{<i]}+U^{-1}_{<i>}U_{<i]})\}\exp\{-\frac{1}{2}\lambda_i^{-1}U_{ii|<i>}\}\prod_{i\in V}\lambda_i^{-\frac{1}{2}\alpha_i}.$$

It is clear that for each λ_i the factor $\exp\{-\frac{1}{2}\lambda_i^{-1}(\beta_{<i]}+U^{-1}_{<i>}U_{<i]})^t U_{<i>}(\beta_{<i]}+U^{-1}_{<i>}U_{<i]})\}$ is maximized at $\beta_{<i]}=-U^{-1}_{<i>}U_{<i]}$. Note also that $\exp\{-\frac{1}{2}\lambda_i^{-1}U_{ii|<i>}\}\prod_{i\in V}\lambda_i^{-\frac{1}{2}\alpha_i}$ corresponds to the distribution $IG(\alpha_i/2-1, U_{ii|<i>}/2)$ and thus its mode is equal to $\dfrac{U_{ii|<i>}}{\alpha_i}$. Combining the above two results the mode of $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$ is given by

$$\times_{i=1}^m\left(\frac{U_{ii|<i>}}{\alpha_i}, -U^{-1}_{<i>}U_{<i]}\right).$$

The following result on the posterior mode of $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$ now follows immediately from the above calculations.

Proposition 7.2. *Let $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n$ be i.i.d. observations from a centered normal distribution parametrized by $\Xi_{\mathcal{G}}$ with prior $\pi^{\Xi_{\mathcal{G}}}_{U,\alpha}$, and let $S=\frac{1}{n}\sum_{i=1}^n \mathbf{Y}_i\mathbf{Y}_i^t$ be the empirical covariance matrix. From Lemma 5.1 the posterior distribution is equal to $\pi^{\Xi_{\mathcal{G}}}_{nS+U,\alpha+n}$ with posterior mode given as follows:*

$$\times_{i=1}^m\left(\frac{(nS+U)_{ii|<i>}}{\alpha_i+n}, -\left(nS_{<i>}+U_{<i>}\right)^{-1}\left(nS_{<i]}+U_{<i]}\right)\right).$$

PART I: DAG Wishart densities for perfect DAGs

8 Induced priors on $\mathrm{P}_{\mathcal{G}}$ and $\mathrm{Q}_{\mathcal{G}}$ for perfect DAGs

The prior $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$ on $\Theta_{\mathcal{G}}$ (the modified Cholesky space) induces a prior on $\mathrm{P}_{\mathcal{G}}$. When $\mathcal{G}$ is a perfect DAG, the induced prior on $\mathrm{P}_{\mathcal{G}}$ can be evaluated in a relatively straightforward manner. Recall from §4.3 that $\mathrm{P}_{\mathcal{G}}$ is the space of positive definite matrices Ω s.t. $\Omega^{-1}\in \mathrm{PD}_{\mathcal{G}}$. As pointed out earlier in §4.3 when $\mathcal{G}$ is a perfect DAG $\mathrm{P}_{\mathcal{G}}$ corresponds to the space of positive definite matrices with zero restrictions according to the decomposable graph $\mathcal{G}^{\mathrm{u}}$. We now provide an expression for the induced prior on $\mathrm{P}_{\mathcal{G}}$ to enable comparisons between our DAG Wishart distributions and other classes of distributions that have been introduced in the literature. Note that the when $\mathcal{G}$ is a perfect DAG the bijective mapping

$$\psi := ((D,L)\mapsto LD^{-1}L^t) : \Theta_{\mathcal{G}}\to \mathrm{P}_{\mathcal{G}} \tag{8.1}$$

is a diffeomorphism between two open subsets of Euclidean space. The lemma below provides the Jacobian required for deriving the induced priors on $\mathrm{P}_{\mathcal{G}}$. Note that for $a\subset V$ the number of elements of a is denoted by $|a|$.

Lemma 8.1. *Let $\mathcal{G}$ be a perfect DAG. Then the Jacobian of the mapping*[8] $\psi = ((D, L) \mapsto LD^{-1}L^t) : \Theta_{\mathcal{G}} \to \mathrm{P}_{\mathcal{G}}$ *is equal to*

$$\prod_{j=1}^{m} D_{jj}^{-(pa_j+2)}. \tag{8.2}$$

A variant of the proof of this lemma can be found in [21, 9]. We nevertheless give a proof in the Appendix for completeness and because the mapping under consideration is slightly different. Furthermore the notation in the proof is important for subsequent sections.

Lemma 8.1 allows us to write the density of the induced prior on $\mathrm{P}_{\mathcal{G}}$ as follows:

$$\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}(\Omega) := z_{\mathcal{G}}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega U)\} \prod_{i=1}^{m} D_{ii}(\Omega)^{-\frac{1}{2}\alpha_i + pa_i + 2}. \tag{8.3}$$

Here $D_{ii} = (\Omega^{-1})_{ii|<i>}$ is considered as a function of Ω.

Recall that if $\mathcal{G}$ is a perfect DAG, then $\mathcal{G}^{\mathrm{u}}$ is decomposable and the sets $\mathrm{P}_{\mathcal{G}}$ and $P_{\mathcal{G}^{\mathrm{u}}}$ are identical. So it is natural to ask whether the distributions we define on $\mathrm{P}_{\mathcal{G}}$ are comparable with other distributions in the literature that are defined on the same space in the decomposable undirected graph setting. We first note that the traditional or classical Wishart distribution [16] on $\mathrm{PD}_m(\mathbb{R})$ is a special case of $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$. In particular the standard Wishart distribution with scale parameter U and degrees of freedom n is a special case of $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ when $\mathcal{G}$ is a perfect DAG with $pa(i) = \{i+1, \ldots, m\}$, and $\alpha_i = n + m - 2i + 3,\ \forall\ 1 \le i \le m$. In this sense we can regard $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ as a generalization of the classical Wishart dsitribution. We also note that the $\mathcal{G}$-Wishart distribution introduced in [21] for undirected graphs, that is the inverse of the hyper-inverse Wishart of [7] which has a one-dimensional shape parameter δ, is also a special case of the richer class $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$. The single shape parameter δ for the $\mathcal{G}$-Wishart is related to the α_i as follows: $\alpha_i = \delta + 2pa_i + 2,\ 1 \le i \le m$.

In the decomposable graph setting, a more general family of distributions on $\mathrm{P}_{\mathcal{G}}$ are the so-called type II Wishart distributions, denoted by $\mathrm{W}_{\mathrm{P}_{\mathcal{G}}}$, introduced in the seminal work of Letac and Massam in [15], and later successfully used by Rajaratnam et al.[19] for high dimensional Bayesian inference for undirected graphs. Recall that perfect DAG models and decomposable undirected graphical models are Markov equivalent, and therefore the corresponding parameter spaces are the same. Hence a careful comparison between the DAG Wishart distributions introduced in this paper and the $\mathrm{W}_{\mathrm{P}_{\mathcal{G}}}$ Wishart distributions introduced in [15] is very important. The multiple shape parameter for the $\mathrm{W}_{\mathrm{P}_{\mathcal{G}}}$ family belongs to a set $\mathcal{B}$, and is fully known only when the graph is homogeneous. In particular the set $\mathcal{B}$ in which the shape parameters lie is not fully characterized when $\mathcal{G}$ is decomposable. In [15] the authors however show that for any perfect ordering $\mathcal{P}$ of a given decomposable graph, there exists a well-describable set $\mathcal{B}_{\mathcal{P}} \subset \mathcal{B}$ over which the corresponding normalizing constants are finite and independent of the scale parameter. Moreover, they conjecture that $\mathcal{B}$

[8] Conventionally, by the Jacobian of a mapping we mean the absolute value of the determinant of the Jacobian matrix.

is indeed the union of $\mathcal{B}_{\mathcal{P}}$ over all perfect ordering of the cliques of the graph $\mathcal{G}$. In an interesting and more involved development (see [4]) we show that for any non-homogeneous decomposable graph $\mathcal{G}$ there is a perfect ordering $\mathcal{P}$ and a perfect DAG, $\mathcal{G}$, associated with this ordering such that the Wishart distribution $\mathrm{W}_{P_{\mathcal{G}}}$ on $\mathcal{B}_{\mathcal{P}}$ is a special case of $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ for a specific choice of α. Under this observation we prove in [4] that the Letac-Massam conjecture is in general not true. When the graph H is homogeneous, we also show that there exists a unique perfect DAG version $\mathcal{G}$ of H such that $\mathrm{W}_{\mathrm{P}_{\mathcal{G}}}$ is a special case of our distribution on $\mathrm{P}_{\mathcal{G}}$. We emphasize here that the analysis of DAGs undertaken in this paper was the primary key to resolving the aforementioned conjecture, which in fact were developed for undirected (or concentration) graph models. The results in [4] which prove that $\mathrm{W}_{\mathrm{P}_{\mathcal{G}}}$, introduced in [15], are a special case of $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ are not the real focus here, are rather involved and a subject of interest in their own right. We simply mention these technical mathematical results for comparison purposes as a detailed study is beyond the scope of this paper. Our focus in this paper rather, is to study the properties of our DAG Wishart distributions with the specific goal of using them for high dimensional Bayesian inference for DAGs.

Another question of interest is the functional form of the induced prior on the space $\mathrm{Q}_{\mathcal{G}}$. A preliminary result is required in order to determine this image measure. Suppose $\mathcal{G}$ is a perfect DAG and $\mathcal{G}^{\mathrm{u}} = (V, E^{\mathrm{u}})$ is the undirected version of $\mathcal{G}$. Grőne *et al.*. [8] prove that for any incomplete matrix Γ in $\mathrm{Q}_{\mathcal{G}}$ there exists a unique $\Sigma(\Gamma)$ in $\mathrm{PD}_{\mathcal{G}}$ such that $\Sigma^E = \Gamma$, where Σ^E is the image of Σ under the projection mapping $\mathrm{proj}_{\mathcal{G}} : \mathrm{S}_m(\mathbb{R}) \to \mathrm{I}_{\mathcal{G}}$. This defines an isomorphism [9] between $\mathrm{Q}_{\mathcal{G}}$ and $\mathrm{P}_{\mathcal{G}}$ via:

$$\varphi := (\Gamma \mapsto \Sigma(\Gamma)^{-1}) : \mathrm{Q}_{\mathcal{G}} \to \mathrm{P}_{\mathcal{G}}$$
$$\varphi^{-1} = (\Omega \mapsto (\Omega^{-1})^E) : \mathrm{P}_{\mathcal{G}} \to \mathrm{Q}_{\mathcal{G}}$$

The matrix $\Sigma(\Gamma)$ is said to be the (positive definite) completion of Γ in $\mathrm{PD}_{\mathcal{G}}$.

The Jacobian of the mapping $\Gamma \mapsto \Sigma(\Gamma)^{-1}$, given in [21], is as follows:

$$\frac{\prod_{s\in\mathscr{S}} |\Gamma_s|^{(|s|+1)\nu(s)}}{\prod_{c\in\mathscr{C}_{\mathcal{G}}} |\Gamma_c|^{|c|+1}}, \quad \Gamma \in \mathrm{Q}_{\mathcal{G}}. \tag{8.4}$$

Consequently, the induced prior on $\mathrm{Q}_{\mathcal{G}}$ is given as

$$\pi^{\mathrm{Q}_{\mathcal{G}}}_{U,\alpha}(\Gamma) \propto \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\}\frac{\prod_{s\in\mathscr{S}} |\Gamma_s|^{(|s|+1)\nu(s)}}{\prod_{c\in\mathscr{C}_{\mathcal{G}}} |\Gamma_c|^{|c|+1}} \prod_{i=1}^{m} D_{ii}(\Gamma)^{-\frac{1}{2}\alpha_i + pa_i + 2}, \quad \Gamma \in \mathrm{Q}_{\mathcal{G}}.$$

Evidently, since $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ is a generalization of the classical Wishart distribution and the G-Wishart, hence $\pi^{\mathrm{Q}_{\mathcal{G}}}_{U,\alpha}$ above is a generalization of the inverse Wishart distribution and the Hyper inverse Wishart (HIW). Furthermore, since the $\mathrm{W}_{\mathrm{P}_{\mathcal{G}}}$ family of Letac-Massam [15] is a special case of $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$, the Inverse of $\mathrm{W}_{\mathrm{P}_{\mathcal{G}}}$ denoted by $\mathrm{IW}_{\mathrm{P}_{\mathcal{G}}}$ is also a special case of $\pi^{\mathrm{Q}_{\mathcal{G}}}_{U,\alpha}$ above.

[9] Furthermore, it also defines a diffeomorphism.

8.1 Closed form expressions for perfect DAGs

We now provide closed form expressions for expected values of Ω and the incomplete positive definite (random) matrix $\Gamma \in \mathrm{Q}_{\mathcal{G}}$, when $\mathcal{G}$ is perfect. The main reason for restricting $\mathcal{G}$ to perfect DAGs is that in this case $\mathrm{P}_{\mathcal{G}}$ and $\mathrm{Q}_{\mathcal{G}}$ are open subsets of the Euclidean space $\mathbb{R}^{|E|}$ and therefore integrations w.r.t. Lebesgue measure are meaningful and the expected values are indeed defined within the parameter spaces.

First note that when $\mathcal{G}$ is perfect the Laplace transform of $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ at $K \in \mathrm{S}_m(\mathbb{R})$ is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{P}_{\mathcal{G}}}(K) &= \int \exp\{-\mathrm{tr}(K\Omega)\}\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}(\Omega)d\Omega \\ &= z^{-1}_{\mathcal{G}}(U,\alpha)\int \exp\{-\frac{1}{2}\mathrm{tr}((2K+U)\Omega)\}\prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i+pa_i+2}d\Omega \\ &= \frac{z_{\mathcal{G}}(2K+U,\alpha)}{z_{\mathcal{G}}(U,\alpha)}.\end{aligned}$$

The following proposition gives the expected value of $\Omega \sim \pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ when $\mathcal{G}$ is perfect.

Proposition 8.1. *Suppose $\mathcal{G}$ is perfect and $\Omega \sim \pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ with $\alpha_i > pa_i + 2$, then*

$$\mathbb{E}[\Omega] = \sum_{i=1}^{m}(\alpha_i - pa_i - 2)\left(U^{-1}_{\leq i\geq}\right)^0 - \sum_{i=1}^{m}(\alpha_i - pa_i - 3)\left(U^{-1}_{<i>}\right)^0.$$

Proof. The proof is given in the Appendix/Supplemental section. □

Our next goal is to determine the expected value of $\Gamma \sim \pi^{\mathrm{Q}_{\mathcal{G}}}_{U,\alpha}$. Note that in essence we can identify $\Gamma \in \mathrm{Q}_{\mathcal{G}}$ with its positive definite completion $\Sigma = \Sigma(\Gamma)$. Under this consideration we show that by a recursive algorithm one can calculate the expected value of Σ.

Proposition 8.2. *Let $\mathcal{G}$ be a perfect DAG and $\Gamma \sim \pi^{\mathrm{Q}_{\mathcal{G}}}_{U,\alpha}$, with $\alpha_i > pa_i+4$. Then the expected value of Γ can be recursively computed in the following steps:*

(i) $\mathbb{E}[\Sigma_{mm}] = \dfrac{U_{mm}}{\alpha_m - 4}$,

(ii) $\mathbb{E}[\Sigma_{<i]}] = -\mathbb{E}[\Sigma_{<i>}]U^{-1}_{<i>}U_{<i]}$,

(iii) $\mathbb{E}(\Sigma_{ii}) = \dfrac{U_{ii|<i>}}{\alpha_i - pa_i - 4} + \mathrm{tr}\left(\mathbb{E}[\Sigma_{<i>}]\left(\dfrac{U_{ii|<i>}U^{-1}_{<i>}}{\alpha_i - pa_i - 4} + U^{-1}_{<i>}U_{<i]}U_{[i>}U^{-1}_{<i>}\right)\right)$, $i = m-1,\ldots,1$.

Proof. The proof is given in the Appendix/Supplemental section. □

Remark 8.1. We note that the recursive expressions in Propostion 8.2 are very similar to the expressions for the expected covariance matrix under the covariance Wishart priors introduced in [9][Corollary 4] for homogeneous covariance graph models. The Markov equivalence of covariance graph models and DAG models for $\mathcal{G}$ homogeneous explains this similarity. We note however that Propostion 8.2 is valid for all perfect graphs in the DAG setting, and is not confined to the restrictive class of homogeneous graph as in the covariance graph setting.

PART II: Hausdorff DAG Wishart densities on curved manifolds

9 Density of $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{G}}}$ w.r.t. Hausdorff measure for an arbitrary DAG

In this section we generalize the prior $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{G}}}$, obtained for a perfect DAG $\mathcal{G}$, to an arbitrary DAG. Recall that $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{G}}}$, the image of the DAG Wishart $\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}$, is a distribution on $\mathrm{P}_{\mathcal{G}}$, the space of concentration matrices in the context of Gaussian DAG models. When $\mathcal{G}$ is a perfect DAG the space $\mathrm{P}_{\mathcal{G}}$ is an open subset of $\mathrm{Z}_{\mathcal{G}} \cong \mathbb{R}^{|E|}$ and therefore $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{G}}}$ has a density w.r.t. Lebesgue measure on $\mathbb{R}^{|E|}$. The functional form of this density was derived in Equation (8.3). When $\mathcal{G}$ is no longer a perfect DAG several complications arise. In particular, the space $\mathrm{P}_{\mathcal{G}}$ has Lebesgue measure zero in any Euclidean vector space containing it. This implies that $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{G}}}$ does not have a density w.r.t. Lebesgue measure. In theory a solution to this problem requires deriving the density of $\pi_{U,\alpha}^{\mathrm{P}_{\mathcal{G}}}$ w.r.t. Hausdorff measure. This section elaborates on this topic in much detail.

9.1 Lebesgue measure of $\mathrm{P}_{\mathcal{G}}$

In this section we undertake a measure theoretic analysis of the space $\mathrm{P}_{\mathcal{G}}$ when $\mathcal{G}$ is not a perfect DAG. First note that Lemma 2.1 implies the following: $\mathrm{P}_{\mathcal{G}} \subset \mathrm{P}_{\mathcal{G}^{\mathrm{m}}} \subset \mathrm{Z}_{\mathcal{G}^{\mathrm{m}}}$. Now let $\mathcal{G} = (V, E)$ be a non-perfect DAG, then $\mathrm{P}_{\mathcal{G}}$ has Lebesgue measure zero in any Euclidean vector space containing it. The next lemma gives a formal proof of this assertion.

Lemma 9.1. *Suppose $\mathcal{G} = (V, E)$ is a non-perfect DAG and $\mathcal{V}$ a Euclidean space containing $\mathrm{P}_{\mathcal{G}}$. Then $\mathcal{V}$ contains $\mathrm{Z}_{\mathcal{G}^{\mathrm{m}}}$. Moreover, $\mathrm{P}_{\mathcal{G}}$ has Lebesgue measure zero in $\mathcal{V}$.*

Proof. For each $(i, j) \in E^{\mathrm{m}}$ with $j \leq i$ let us define the elementary symmetric matrix $\tilde{E}^{(ij)} \in \mathrm{S}_m(\mathbb{R})$ as follows:

$$\tilde{E}^{(ij)}_{uv} = \begin{cases} 1 & \text{if } \{u, v\} = \{i, j\}, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the set of $\tilde{E}^{(ij)}$ forms a basis of $\mathrm{Z}_{\mathcal{G}^{\mathrm{m}}}$. It is clear that $\mathcal{V}$ contains $\mathrm{Z}_{\mathcal{G}} \supset \{\tilde{E}^{(ij)} : (i, j) \in E\}$. Hence it suffices to prove that $\mathcal{V}$ contains the rest of $\tilde{E}^{(ij)}$. Now let (i, j) be in $E^{\mathrm{m}} \setminus E$ with $i > j$. This implies that there exists $k < j < i$ such that $i \to k \leftarrow j$. We define the lower triangular matrix $L^{(ij)} \in \mathcal{L}_{\mathcal{G}}$ as follows:

$$L^{(ij)}_{uv} = \begin{cases} 1 & \text{if } (u, v) = (i, k), \\ 1 & \text{if } (u, v) = (j, k), \\ 1 & \text{if } u = v, \\ 0 & \text{otherwise.} \end{cases}$$

Then one can easily check that $\mathrm{P}_{\mathcal{G}} \ni L^{(ij)}(L^{(ij)})^t = T + 2\tilde{E}^{(ij)}$, for some $T \in \mathrm{Z}_{\mathcal{G}}$. This shows that $\tilde{E}^{(ij)} \in \mathcal{V}$. Hence $\mathrm{P}_{\mathcal{G}} \subset V \Rightarrow \mathrm{Z}_{\mathcal{G}^{\mathrm{m}}} \subset \mathcal{V}$, thus $\mathrm{P}_{\mathcal{G}} \subset \mathrm{Z}_{\mathcal{G}^{\mathrm{m}}} \subset \mathcal{V}$.

Now note that $\mathrm{P}_{\mathcal{G}}$ is a manifold of dimension $|E|$ diffeomorphic to $\Theta_{\mathcal{G}}$, which in turn is an open subset of Euclidean space of dimension $|E|$. Furthermore, recall that the dimension

of $Z_{\mathcal{G}^m} = |E^m|$ and is therefore strictly larger than the $|E|$. So any Euclidean space that contains $P_{\mathcal{G}}$ has dimension strictly larger than $|E|$. Hence $P_{\mathcal{G}}$ has Lebesgue measure zero in any Euclidean vector space containing it. □

Consequently, Lemma 9.1 implies that if $\mathcal{G}$ is non-perfect then $\pi_{U,\alpha}^{P_{\mathcal{G}}}$ has no density w.r.t. Lebesgue measure.

9.2 The density of $\pi_{U,\alpha}^{P_{\mathcal{G}}}$ w.r.t. Hausdorff measure

We now proceed to derive the density of $\pi_{U,\alpha}^{P_{\mathcal{G}}}$ w.r.t. Hausdorff measure [10]. Let $\Delta_{\mathcal{G}}$ denote the set of (D, L) such that $D \in R^{m\times m}$ is a diagonal matrix and $L \in \mathcal{L}_{\mathcal{G}}$. It is immediate that $\Delta_{\mathcal{G}}$ is a real linear space of dimension $|E|$ with the following scaler product and sum operation, respectively.

1. $\lambda(D, L) := (\lambda D, \lambda L), \quad \forall \lambda \in R$;
2. $(D', L') + (D'', L'') = (D, L)$, where $D = (D' + D'')$, and L is a lower triangular matrix with $L_{ij} = L'_{ij} + L''_{ij}$ if $i \neq j$ and $L_{ii} = 1$.

One can easily check that $\Theta_{\mathcal{G}}$ is an open subset of $\Delta_{\mathcal{G}}$. Now since $P_{\mathcal{G}}$ is a subset of Euclidian space Z_{G^m} we have $\psi : \Theta_{\mathcal{G}} :\to Z_{G^m}$ satisfies the conditions of Theorem 19.3 in [2]. Hence we can proceed to obtain the density of $\pi_{U,\alpha}^{P_{\mathcal{G}}}$ w.r.t. the $|E|$-dimensional Hausdorff measure on $Z_{\mathcal{G}^m}$. To obtain an explicit expression for $J(\psi(D, L))$ we first need to compute the matrix of partial derivatives $\frac{\partial \psi_{kl}}{\partial D_{ii}}$ and $\frac{\partial \psi_{kl}}{\partial L_{ij}}$. We order the coordinates of $\Delta_{\mathcal{G}}$ as follows: D_{11}, L_{21} if $(2,1) \in E, D_{22}, L_{31}$ if $(3,1) \in E, L_{32}$ if $(3,2) \in E, \ldots, D_{(m-1)(m-1)}, L_{ml}, l = 1, \ldots (m-1)$ if $(m, l) \in E, D_{mm}$. Likewise, we order the coordinates of $Z_{\mathcal{G}^m} \cong \mathbb{R}^{|E|} \times \mathbb{R}^{|\mathscr{I}|}$, where $\mathscr{I} := E^m \setminus E$, by ordering first the positions $(k, l) \in E$ as above, in their entirety, and then we order the positions $(k, l) \in \mathscr{I}$ according to their lexicographical order. Note that the latter positions correspond to immoralities. These partial derivatives can be computed as follows:

$$\frac{\partial (LD^{-1}L^t)_{kl}}{\partial D_{ii}} = -D_{ii}^{-2} L_{ki} L_{li} \tag{9.1}$$

$$\frac{\partial (LD^{-1}L^t)_{kl}}{\partial L_{ij}} = \delta_{ik} D_{jj}^{-1} L_{lj} + \delta_{il} D_{jj}^{-1} L_{kj}, \tag{9.2}$$

where δ_{uv} is the Kronecker delta function. Using Equations (9.1) and (9.2) we partition the Jacobian matrix $D\psi(D, L)$, considered as a mapping from $\mathbb{R}^{|E|}$ to $\mathbb{R}^{|E|} \times \mathbb{R}^{|\mathscr{I}|}$, into two blocks of matrices $A_\psi := D\psi(L, D)_{EE}$ of size $|E| \times |E|$ and $C_\psi := D\psi(D, L)_{\mathscr{I}E}$ of size $|\mathscr{I}| \times |E|$, respectively. The matrix A_ψ is the same as the Jacobian matrix from Lemma 8.1, and C_ψ is the last $|\mathscr{I}|$-th rows of the Jacobian matrix $D\psi(D, L)$, with each row of C_ψ being the partial

[10] The reader is referred to [2, Section 19] for more details on this topic.

derivatives obtained by Equations (9.1) and (9.2) for $(k,l) \in \mathscr{I}$ and $(i,j) \in E$. Finally, we can calculate the Jacobian of ψ as follows:

$$\begin{aligned}
\mathbf{J}\psi(D,L) &= \det\left(\begin{pmatrix} A_\psi^t & \vdots & C_\psi^t \end{pmatrix}\begin{pmatrix} A_\psi \\ \cdots \\ C_\psi \end{pmatrix}\right)^{\frac{1}{2}} \\
&= \sqrt{\det(A_\psi^t A_\psi + C_\psi^t C_\psi)} \\
&= |\det(A_\psi)| \sqrt{\det(I + A_\psi^{-t} C_\psi^t C_\psi A_\psi^{-1})} \\
&= \prod_{j=1}^{m} D_{jj}^{-(pa_j+2)} \sqrt{\det(I + A_\psi^{-t} C_\psi^t C_\psi A_\psi^{-1})}.
\end{aligned}$$

Therefore we have proved the following.

Theorem 9.1. *Let A_ψ, C_ψ be defined as the block matrices in partitioning of the (Hausdorff) Jacobian matrix of ψ above. Then the density of $\pi_{U,\alpha}^{\mathrm{P}_\mathcal{G}}$ w.r.t. Hausdorff measure $\mathcal{H}^{|E|}$ on $\mathrm{Z}_{\mathcal{G}^{\mathrm{m}}}$ is given by*

$$z_\mathcal{G}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega U)\} \prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i + pa_i + 2} \det(I + A_\psi^{-t} C_\psi^t C_\psi A_\psi^{-1})^{-\frac{1}{2}}. \tag{9.3}$$

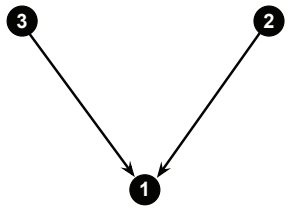

Figure 1: Wishart density w.r.t. a Hausdorff measure

Example 9.1. Consider DAG $\mathcal{G}$ given in Figure 1. The Jacobian matrix corresponding to Equation (9.1) and Equation (9.2) are given as follows:

$$M_\psi = \begin{pmatrix}
-D_{11}^{-2} & 0 & 0 & 0 & 0 \\
-L_{21}D_{11}^{2} & D_{11}^{-1} & 0 & 0 & 0 \\
-L_{21}^2 D_{11}^{-2} & 2L_{21}D_{11}^{-1} & -D_{22}^{-2} & 0 & 0 \\
-L_{31}D_{11}^{-2} & 0 & 0 & D_{11}^{-1} & 0 \\
-L_{31}^2 D_{11}^{-2} & 0 & 0 & 2L_{31}D_{11}^{-1} & 0 \\
-L_{21}L_{31}D_{11}^{-2} & L_{31}D_{11}^{-1} & 0 & L_{21}D_{11}^{-1} & -D_{33}^{-2}
\end{pmatrix}$$

By computing $\det(M_\psi^t M_\psi)$ we obtain

$$\mathbf{J}\psi(D,L) = D_{11}^{-4} D_{22}^{-2} D_{33}^{-2} \left(L_{31}^4 + 4L_{31}^2 + 1\right)^{1/2}.$$

Thus the density of $\pi_{U,\alpha}^{\mathrm{P}_\mathcal{G}}$ w.r.t. $\mathcal{H}^5$ on $\mathbb{R}^6$ is given by

$$z_\mathcal{G}(U,\alpha)^{-1}\exp\{-\frac{1}{2}\mathrm{tr}(\Omega U)\}D_{11}^{-\frac{\alpha_1}{2}+4}D_{22}^{-\frac{\alpha_2}{2}+2}D_{33}^{-\frac{\alpha_3}{2}+2}\left(L_{31}^4+4L_{31}^2+1\right)^{-1/2},$$

where D_{ii} and L_{ij} are considered as functions of Ω.

PART III: DAG Wishart densities on incomplete covariance spaces

10 The DAG Wishart distribution on the space of incomplete concentration matrices

The previous section demonstrated the difficulty of working with $\pi_{U,\alpha}^{\mathrm{P}_\mathcal{G}}$ for a general DAG $\mathcal{G}$: first, the density does not exist w.r.t. Lebesgue measure but only w.r.t to Hausdorff measure, and secondly even w.r.t. Hausdorff measure the density $\pi_{U,\alpha}^{\mathrm{P}_\mathcal{G}}$ is not easily computable. It is not immediately clear how to overcome this obstacle. We remind the reader that this problem does not occcur when $\mathcal{G}$ is restricted to the perfect/decomposable as treated in the work of [15, 9, 7, 19]. To avoid the complexity inherent in the nature of $\mathrm{P}_\mathcal{G}$, we propose an approach to work with only the functionally independent elements of $\mathrm{P}_\mathcal{G}$ and consequently demonstrate that this can lead to fruitful results. More precisely, we consider the projection of $\mathrm{P}_\mathcal{G}$ onto the space of incomplete symmetric matrices where the specified entries are in positions determined by the edge set of $\mathcal{G}$, i.e., along the edges. To this end we shall first demonstrate that $\mathrm{P}_\mathcal{G}$ can be easily identified with a new space $\mathrm{R}_\mathcal{G}$ through a simple isomorphism, where the latter is Euclidean from a topological perspective and on which the standard Lebesgue measure is defined.

10.1 The space of incomplete concentration matrices

First recall the definitions of the sets $\mathrm{Z}_\mathcal{G}$ and $\mathrm{I}_\mathcal{G}$ from §4.3 and that $\mathrm{P}_\mathcal{G}$ is the space of concentration matrices corresponding to the Gaussian DAG model $\mathcal{N}(\mathcal{G})$. There is a natural injection $(\Gamma \mapsto (\Gamma)^0) : \mathrm{I}_\mathcal{G} \to \mathrm{Z}_\mathcal{G}$, where $(\Gamma)^0$, as defined earlier, "fills" or "completes" the unspecified positions with zeros to obtain a full matrix in $\mathbb{R}^{V\times V}$. Note that for each clique c of $\mathcal{G}$ the restriction of Γ on c, denoted by Γ_c is a full matrix and, moreover, Γ is uniquely determined by the blocks of matrices $(\Gamma_c : c \in \mathscr{C}_\mathcal{G})$.

Definition 10.1. Suppose $\mathcal{K} \subset \mathbb{R}^{m\times m}$ and Γ is a $\mathcal{G}$-incomplete matrix in $\mathrm{I}_\mathcal{G}$, then we say Γ can be completed in $\mathcal{K}$ if there exists a matrix $\mathrm{A} \in \mathcal{K}$ such that $\mathrm{A}_c = \Gamma_c$ for each clique $c \in \mathscr{C}_\mathcal{G}$.

Remark 10.1. If $\mathcal{K}$ is the set of positive definite matrices $\mathrm{PD}_m(\mathbb{R})$ then the completion defined above reduces to the standard definition of positive definite completion [8]. We note however that in what ensues below, the positive definite completion refers to completion of partially positive concentration and covariance matrices that correspond to DAGs vs. those that correspond to undirected graphs as in [8]. Note also that a necessary condition

for a positive definite completion of an incomplete matrix is that it belongs to $Q_{\mathcal{G}}$. This requirement is simply derived from the condition that the principal minors of positive definite matrices are positive definite. Thus we shall henceforth only focus on completion of partially positive definite matrices over $\mathcal{G}$, i.e., those elements in $Q_{\mathcal{G}}$ as compared to the larger class $I_{\mathcal{G}}$.

Proposition 10.1. *Let Υ be a $\mathcal{G}$-incomplete matrix in $I_{\mathcal{G}}$. Then:*

1. *Almost everywhere (w.r.t. Lebesgue measure on $I_{\mathcal{G}}$), there exist a lower triangular matrix $L \in \mathcal{L}_{\mathcal{G}}$ and a diagonal matrix $\Lambda \in R^{m\times m}$ such that $\widehat{\Upsilon} := L\Lambda L^t$ is a completion of Υ.*

2. *The completion algorithm to construct Λ and L are given as follows:*

   i) *Set $L_{ij} = 0$ for each $(i, j) \notin E$.*

   ii) *Set $\Lambda_{11} = \Gamma_{11}$, $L_{i1} = \Lambda_{11}^{-1}\Gamma_{i1}$ for each $i \in \mathrm{pa}(1)$ and set $j = 1$.*

   iii) *If $j < p$, then set $j = j + 1$ and proceed to step v), otherwise L and Λ are constructed such that they satisfy the condition in part (a).*

   v) *Set $\Lambda_{jj} = \Gamma_{jj} - \sum_{k=1}^{j-1} \Lambda_{kk} L_{jk}^2$ and proceed to the next step.*

   vi) *For each $i \in \mathrm{pa}(j)$ if $\Lambda_{jj} \neq 0$, then set $L_{ij} = \Lambda_{jj}^{-1}(\Gamma_{ij} - \sum_{k=1}^{j-1} \Lambda_{kk} L_{ik} L_{jk})$, and return to step iii). If $\Lambda_{jj} = 0$, then no completion of Υ exists that satisfies the condition in part (a). Consequently, Υ cannot also be completed in $P_{\mathcal{G}}$.*

3. *The matrix $\widehat{\Upsilon}$ is the unique positive definite completion of Υ in $P_{\mathcal{G}}$ iff the diagonal entries of Λ are all strictly positive.*

Proof. The proof is found in [5]. □

Remark 10.2. Note first that in Proposition 10.1 the algorithm itself determines if Υ can be completed in $P_{\mathcal{G}}$. Furthermore, the process described in Proposition 10.1 succeeds to complete Υ in the space of symmetric matrices $S_m(\mathbb{R})$ as long as $\lambda_{jj} \neq 0$. However, the completion is in $P_{\mathcal{G}}$ iff Λ_{jj} are all strictly positive.

Example 10.1. Let $\mathcal{G}$ be the DAG given by Figure 10.1.
(a) Let Υ_1 be the $\mathcal{G}$-incomplete matrix

$$\Upsilon_1 = \begin{pmatrix} 4 & 8 & 8 & ? \\ 8 & 19 & ? & 9 \\ 8 & ? & 18 & 6 \\ ? & 9 & 6 & 44 \end{pmatrix}.$$

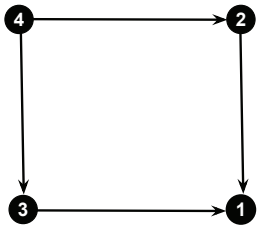

Figure 2: Completion in $\mathrm{P}_{\mathcal{G}}$

Applying the algorithm described in Proposition 10.1 we obtain:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & 3 & 3 & 1 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & -1 \end{pmatrix}, \quad \widehat{\Upsilon}_1 = \begin{pmatrix} 4 & 8 & 8 & 0 \\ 8 & 19 & 16 & 9 \\ 8 & 16 & 18 & 6 \\ 0 & 9 & 6 & 44 \end{pmatrix}.$$

The negative element in Λ demonstrates that Υ_1 cannot be completed in $\mathrm{P}_{\mathcal{G}}$.

(*b*) Let Υ_2 be the $\mathcal{G}$-incomplete matrix

$$\Upsilon_2 = \begin{pmatrix} 1 & 1 & 2 & ? \\ 1 & 3 & ? & -2 \\ 2 & ? & 5 & 1 \\ ? & -2 & 1 & 4 \end{pmatrix}.$$

Applying the algorithm in Proposition 10.1 once more, we obtain

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 0 & -1 & 1 & 1 \end{pmatrix}, \quad \Lambda = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \widehat{\Upsilon}_2 = \begin{pmatrix} 1 & 1 & 2 & 0 \\ 1 & 3 & 2 & -2 \\ 2 & 2 & 5 & 1 \\ 0 & -2 & 1 & 4 \end{pmatrix}.$$

Proposition 10.1 guarantees that the above yields the unique positive definite completion of Υ_2 in $\mathrm{P}_{\mathcal{G}}$.

The following corollary is an immediate consequence of Proposition 10.1.

Corollary 10.1. *Let* $\mathrm{R}_{\mathcal{G}}$ *denote the set of* $\Upsilon \in \mathrm{Q}_{\mathcal{G}}$ *that can be completed in* $\mathrm{P}_{\mathcal{G}}$. *Then the mapping* $(\Omega \mapsto \Omega^E) : P_{\mathcal{G}} \to \mathrm{R}_{\mathcal{G}}$ *is a bijection with the inverse mapping* $\Upsilon \mapsto \widehat{\Upsilon}$.

In essence Corollary 10.1 identifies our concentration matrix space $\mathrm{P}_{\mathcal{G}}$ with another space $R_{\mathcal{G}}$ through a bijection. The space $R_{\mathcal{G}}$ is an open subset of Euclidean space $\mathbb{R}^{|E|}$ since $R_{\mathcal{G}}$ is homeomorphic to $\Theta_{\mathcal{G}}$. We shall henceforth refer to $\mathrm{R}_{\mathcal{G}}$ as the space of incomplete concentration matrices over $\mathcal{G}$.

10.2 The Wishart distribution on $\mathrm{R}_{\mathcal{G}}$

Recall that given a matrix $\Omega \in \mathrm{P}_{\mathcal{G}}$, the $\mathcal{G}$-incomplete matrix Ω^E contains all the functionally independent entries of Ω and, by Proposition 10.1, one can always recover the remaining entries of Ω in polynomial time. Let $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ denote the image of $\pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ under the mapping $\Omega \mapsto \Omega^E$. Since $\mathrm{R}_{\mathcal{G}}$ is an open subset of Euclidean space $\mathbb{R}^{|E|}$ this distribution has a density w.r.t. Lebesgue measure on $\mathrm{R}_{\mathcal{G}}$. Hence $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ can be considered as the DAG Wishart distribution on $\mathrm{R}_{\mathcal{G}}$ in both a natural and practical sense. We now proceed to state general results corresponding to our DAG Wishart distributions on the space of incomplete concentration matrices for *ALL* DAGs, and not just perfect DAGs as given in Section 8.1.

Theorem 10.1. *Let $\Omega \sim \pi^{\mathrm{P}_{\mathcal{G}}}_{U,\alpha}$ and let $\Upsilon = \mathrm{proj}(\Omega) = \Omega^E$, then*

1. *The density of $\Upsilon \sim \pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ w.r.t. the standard Lebesgue measure on $\mathrm{R}_{\mathcal{G}}$ is given by*

$$z_{\mathcal{G}}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\widehat{\Upsilon}U)\} \prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i + pa_i + 2},$$

   where $D_{ii} = \left(\widehat{\Upsilon}\right)^{-1}_{ii|<i>}$.

2. *The Laplace transform of $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ at K^E, where $K > 0$, is given by $\mathcal{L}_{R_{\mathcal{G}}}(K^E) = \dfrac{z_{\mathcal{G}}(2K+U,\alpha)}{z_{\mathcal{G}}(U,\alpha)}$.*

3. $\mathbb{E}[\Upsilon] = \mathrm{proj}_{\mathcal{G}}\left(\sum_{j=1}^{m}(\alpha_j - pa_j - 2)\left(U^{-1}_{\leq j \geq}\right)^0 - \sum_{j=1}^{m}(\alpha_j - pa_j - 3)\left(U^{-1}_{<j>}\right)^0\right)$.

Proof. The distribution $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ is the image of $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$ under the mapping $(D,L) \mapsto (LD^{-1}L^t)^E$. Thus it suffices to compute the Jacobian of this mapping. One can readily check that the Jacobian is equal to the Jacobian of ψ in Equation (8.2). Similar calculations as in Section 8.1 yields the Laplace transform and expected value of $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$. □

11 The DAG inverse Wishart on the space of incomplete covariance matrices

A natural question that follows from the last section is to determine the image of $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$ on the space of covariance matrices as denoted by $\mathrm{PD}_{\mathcal{G}}$. This measure is the induced prior on the space of covariance matrices that correspond to a Gaussian DAG model. In this section we first identify a subset of the space of $\mathcal{G}$-incomplete matrices, called $\mathrm{S}_{\mathcal{G}}$, that can be identified with the space of covariance matrices $\mathrm{PD}_{\mathcal{G}}$. Thereafter we shall define the DAG inverse Wishart distribution on the space of incomplete covariance matrices $\mathrm{S}_{\mathcal{G}}$.

11.1 The space of incomplete covariance matrices

Recall that $\mathrm{PD}_{\mathcal{G}}$ is the set of positive definite matrices Σ in $\mathrm{PD}_m(\mathbb{R})$ such that $\mathrm{N}_m(0,\Sigma)$ belongs to $\mathscr{N}(\mathcal{G})$. Equivalently, if Σ is a positive definite matrix, then

$$\Sigma \in \mathrm{PD}_{\mathcal{G}} \quad \text{if and only if} \quad \Sigma_{[i\not>} = \Sigma_{[i>}\Sigma^{-1}_{<i>}\Sigma_{<i\not>}, \text{ for each } i \in V. \tag{11.1}$$

Using this characterization allows us to identify $\mathrm{PD}_{\mathcal{G}}$ with the functionally independent elements of Σ. The following result is a key ingredient in this identification.

Proposition 11.1. *Let $\Gamma \in \mathrm{Q}_{\mathcal{G}}$, then*

1. *There exists a completion process of polynomial complexity that can determine whether Γ can be completed in $\mathrm{PD}_{\mathcal{G}}$;*
2. *If a completion exists, this completion is unique and can be determined constructively using the following process:*
   - *i) Set $\Sigma_{ij} = \Gamma_{ij}$ for each $(i,j) \in E$ and set $j = m$.*
   - *ii) If $j > 1$, then set $j = j - 1$ and proceed to the next step, otherwise Σ is successfully completed.*
   - *iii) If $\Sigma_{\leq j\geq} \succ 0$, then proceed[11] to the next step, otherwise the completion in $\mathrm{PD}_{\mathcal{G}}$ does not exist.*
   - *iv) If $\Sigma_{\not< j]}$ is non-empty, then set $\Sigma_{\not< j]} = \Sigma_{\not< j>}\Sigma^{-1}_{<j>}\Sigma_{<j]}$, $\Sigma_{[j\not>} = \Sigma^t_{\not< j]}$ and return to step (2).*

Proof. The proof is found in [5]. □

Remark 11.1. Note that in Proposition 11.1, once more, the algorithm itself determines if Γ can be completed in $\mathrm{PD}_{\mathcal{G}}$. It is clear from Step *(iii)* above that the necessary and sufficient condition for the positive definite completion to exist is for the covariance sub-matrix of the family of each node j to be positive definite, i.e., $\Sigma_{\leq j\geq} \succ 0$ and not just $\Sigma_{<j>} \succ 0$. Furthermore, unlike Proposition 10.1, Proposition 11.1 can terminate midway.

Now define $\mathrm{S}_{\mathcal{G}}$ as the set of all $\Gamma \in \mathrm{Q}_{\mathcal{G}}$ which can be completed in $\mathrm{PD}_{\mathcal{G}}$. The next corollary formalizes the fact that $\mathrm{S}_{\mathcal{G}}$ can be identified with $\mathrm{PD}_{\mathcal{G}}$.

Corollary 11.1. *Let $\mathcal{G} = (V,E)$ be a DAG, then the mapping $(\Sigma \mapsto \Sigma^E) : \mathrm{PD}_{\mathcal{G}} \to \mathrm{S}_{\mathcal{G}}$ is a bijection with inverse mapping $\Gamma \mapsto \Sigma(\Gamma)$, where $\Sigma(\Gamma)$ is constructed according to Proposition 11.1.*

Proof. The proof is immediate from Proposition 11.1 above. □

Remark 11.2. Note that when $\mathcal{G}$ is perfect $\mathrm{PD}_{\mathcal{G}}$ is identical to $\mathrm{PD}_{\mathcal{G}^{\mathrm{u}}}$. Thus by the completion result in Grone et al. [8], when $\mathcal{G}$ is perfect every partial positive definite matrix in $\mathrm{Q}_{\mathcal{G}}$ can be completed in $\mathrm{PD}_{\mathcal{G}}$. Hence for $\mathcal{G}$ perfect $\mathrm{S}_{\mathcal{G}}$ and $\mathrm{Q}_{\mathcal{G}}$ are identical.

[11] Note that for each j, the submatrix $\Sigma_{\leq j\geq}$ is fully determined by step *(ii)*

We now proceed to illustrate the result of the completion process on two examples.

Example 11.1. (*a*) Consider the DAG $\mathcal{G}$ given in Figure 3(a) and let Γ_1 be the partial positive $\mathcal{G}$-incomplete matrix given as follows:

$$\Gamma_1 = \begin{pmatrix} 1 & 0.9 & ? & -0.9 \\ 0.9 & 1 & 0.9 & ? \\ ? & 1 & 1 & 0.9 \\ -0.9 & ? & 0.9 & 1 \end{pmatrix}.$$

Applying the completion process yields the following results: In step *(iv)* for j=2 we obtain $\Sigma_{42} = 0.9$. From this, in step *(iii)* for $j = 1$ we obtain

$$\Sigma_{\leq 1 \geq} = \begin{pmatrix} 1 & 0.9 & -0.9 \\ 0.9 & 1 & 0.9 \\ -0.9 & 0.9 & 1 \end{pmatrix},$$

which is not a positive definite matrix. Hence the completion process demonstrates that no completion of Γ_1 exists in $\mathrm{PD}_{\mathcal{G}}$.

(a) (b)

Figure 3: Completion in $\mathrm{PD}_{\mathcal{G}}$

(*b*) Consider the DAG $\mathcal{G}$ as given in Figure 3(b) and let Γ be the $\mathcal{G}$-incomplete matrix given by

$$\Gamma = \begin{pmatrix} 3\frac{1}{3} & -1 & \frac{1}{3} & ? & ? & ? \\ -1 & 6 & ? & -5\frac{1}{2} & ? & ? \\ \frac{1}{3} & ? & 22\frac{1}{3} & -11 & ? & ? \\ ? & -5\frac{1}{2} & -11 & 5\frac{1}{2} & -2 & -1 \\ ? & ? & ? & -2 & 1 & ? \\ ? & ? & ? & -1 & ? & 1 \end{pmatrix}.$$

Then by applying the completion process in Proposition 11.1 we obtain

$$\Sigma = \begin{pmatrix} 3\frac{1}{3} & -1 & \frac{1}{3} & 0 & 0 & 0 \\ -1 & 6 & 11 & -5\frac{1}{2} & 2 & 1 \\ \frac{1}{3} & 11 & 22\frac{1}{3} & -11 & 4 & 2 \\ 0 & -5\frac{1}{2} & -11 & 5\frac{1}{2} & -2 & -1 \\ 0 & 2 & 4 & -2 & 1 & 0 \\ 0 & 1 & 2 & -1 & 0 & 1 \end{pmatrix}.$$

From the completion process it is easily verified the matrix Σ is a positive definite matrix. Moreover, Proposition 11.1 guarantees that it is the unique completion of Γ in $\mathrm{PD}_{\mathcal{G}}$.

11.2 The DAG inverse Wishart distribution on $\mathrm{S}_{\mathcal{G}}$

Let $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$ denote the image of $\pi^{\Theta_{\mathcal{G}}}_{U,\alpha}$ under the mapping $(D, L) \mapsto (L^{-t}DL^{t})^{E} : \Theta_{\mathcal{G}} \to \mathrm{S}_{\mathcal{G}}$. Similar to our interpretation of $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ we will consider $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$ as the inverse Wishart distribution for the DAG $\mathcal{G}$. Next we proceed to derive the density of this distribution w.r.t Lebesgue measure. To this end, recall the following notion from [20].

Definition 11.1. Let T be a symmetric matrix in $\mathbb{R}^{V\times V}$, where as usual $V = \{1, \ldots, m\}$. The Isserlis matrix of T, denoted by $\mathrm{Iss}(T)$, is the symmetric matrix indexed by the set $\mathscr{W} = \{(i, j) : i, j \in V, i \geq j\}$ with entries

$$Iss(T)_{ij,kl} = T_{ik}T_{jl} + T_{il}T_{jk}, \quad (i, j), (k, l) \in \mathscr{W}.$$

We also note the following properties of $Iss(T)$:
(i) $\det(Iss(T)) = 2^{m}\det(T)^{m+1}$,
(ii) $Iss(T)$ is invertible if and only if T is invertible.

Now for a subset $\mathscr{U}$ of $\mathscr{W}$ define $T^{\mathscr{U}}$ as the incomplete symmetric matrix the entries of which are specified as $T^{\mathscr{U}}_{ij} = T^{\mathscr{U}}_{ji} = T_{ij}$ for each $(i, j) \in \mathscr{U}$. If in addition, T is invertible, then we shall denote $(T^{-1})^{\mathscr{U}}$ by $T^{-\mathscr{U}}$. We caution the reader that the notation $T^{\mathscr{U}}$ differs from T^{a} where $a \subset V$. In the former, $\mathscr{U} \subset \mathscr{W} \subset V \times V$, whereas the latter refers to $(T^{-1})_{a}$, i.e., $a \subset V$. With this notation in hand we now proceed to state the functional form of the the DAG inverse Wishart distribution $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$.

Proposition 11.2. *Let $\mathcal{G} = (V, E)$ be an arbitrary DAG and let $\Sigma \sim \pi^{\mathrm{PD}_{\mathcal{G}}}_{U,\alpha}$. Now let $\Gamma = \mathrm{proj}(\Sigma) = \Sigma^{E}$, then the density of $\Gamma \sim \pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$ w.r.t. Lebesgue measure is given by*

$$\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}(\Gamma) = 2^{m}z_{\mathcal{G}}(U, \alpha)^{-1}\exp\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\}\prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i + pa_i + 2}\det(\mathrm{Iss}(\Sigma)_{E|\mathscr{I}})^{-1}, \qquad (11.2)$$

where $\mathscr{I} = \mathscr{V} \setminus E$, and where $\mathscr{V}$ is the edge set of $\mathcal{G}^{\mathrm{m}}$, with the convention that if $(i, j) \in \mathscr{V}$, then $i \geq j$, and $D_{ii} = \Sigma(\Gamma)_{ii|<i>}$.

Proof. First note that functional form of $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$ can be obtained as the image of $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ under the mapping $\Upsilon \mapsto \mathrm{proj}_{\mathcal{G}}(\widehat{\Upsilon})^{-1} : \mathrm{R}_{\mathcal{G}} \to \mathrm{S}_{\mathcal{G}}$. Let us denote the inverse of this mapping symbolically by $\Sigma^{E} \mapsto \Sigma^{-E}$. We now proceed to evaluate the Jacobian of this mapping. Following the notation in [20] let $\mathrm{M}^{+}_{*}(\mathcal{G}^{\mathrm{m}})$ denote the set of $\Sigma^{\mathscr{V}}$ such that $\Sigma \in \mathrm{PD}_{\mathcal{G}^{m}}$, and let $\mathrm{M}^{+}(\mathcal{G}^{\mathrm{m}})$ denote the the set of $\Sigma^{-\mathscr{V}}$ such that $\Sigma \in \mathrm{PD}_{\mathcal{G}^{\mathrm{m}}}$. By [20, Equation (11)] the derivative of the mapping $\Sigma^{\mathscr{V}} \mapsto \Sigma^{-\mathscr{V}}$ is given by

$$\frac{\partial\Sigma^{-\mathscr{V}}}{\partial\Sigma^{\mathscr{V}}} = -Iss(\Sigma^{-1})_{\mathscr{V}|\overline{\mathscr{V}}}Iss(I)^{\mathscr{V}|\overline{\mathscr{V}}}, \qquad (11.3)$$

where $\overline{\mathscr{V}} = \mathscr{W} \setminus \mathscr{V}$. Since any distribution that obeys the directed Markov property w.r.t. $\mathcal{G}$ will obey the Markov property w.r.t. $\mathcal{G}^{\mathrm{m}}$ (see Lemma 2.1), hence $\Sigma \in \mathrm{PD}_{\mathcal{G}}$ implies that $\Sigma \in \mathrm{PD}_{\mathcal{G}^{\mathrm{m}}}$. Therefore, the mapping $\Sigma^E \mapsto \Sigma^{-E}$ is the restriction of the mapping $\Sigma^{\mathscr{V}} \mapsto \Sigma^{-\mathscr{V}}$ to $\mathrm{S}_{\mathcal{G}}$. Thus by using Equation (11.3) above we obtain

$$\begin{aligned}\frac{\partial \Sigma^{-E}}{\partial \Sigma^{E}} &= \left(\mathrm{Iss}(\Sigma^{-1})_{\mathscr{V}|\overline{\mathscr{V}}} Iss(I)^{\mathscr{V}|\overline{\mathscr{V}}}\right)_{EE} \\ &= \mathrm{Iss}(\Sigma^{-1})_{E|\overline{\mathscr{V}}} Iss(I)^{E|\overline{\mathscr{V}}} \\ &= \mathrm{Iss}(\Sigma^{-1})_{E|\overline{\mathscr{V}}} Iss(I)^{E},\end{aligned}$$

where the last two steps follow from the fact that $Iss(I)$ is a diagonal matrix and that $Iss(I)^{E|\overline{\mathscr{V}}} = Iss(I)^{E}$. By using Equation (2.1) in [**?**] and Equation (9) in [20] respectively, we can write

$$\begin{aligned}\mathrm{Iss}(\Sigma^{-1})_{E|\overline{\mathscr{V}}} &= \left(\mathrm{Iss}(\Sigma^{-1})^{E|\mathscr{I}}\right)^{-1} \\ &= \left(Iss(I)^{E}\mathrm{Iss}(\Sigma)_{E|\mathscr{I}} Iss(I)^{E}\right)^{-1}.\end{aligned}$$

Hence the Jacobian of the mapping $\Sigma \mapsto \Sigma^{-E}$ is equal to $2^m \det(\mathrm{Iss}(\Sigma)_{E|\mathscr{I}})^{-1}$. The functional form of $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}(\Gamma)$ now follows from a change of measure calculation. □

Remark 11.3. Note that for calculating the Jacobian term in the density $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}(\Gamma)$ above, one only needs to evaluate $\Sigma(\Gamma)^{\mathscr{V}}$, that is the completion of Γ in $\mathrm{PD}_{\mathcal{G}}$, restricted to the entries that correspond to $\mathcal{G}^m$.

We now proceed to illustrate the proposition above on a concrete example.

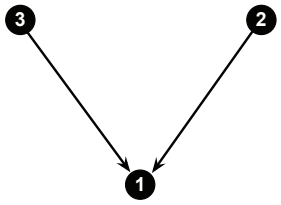

Figure 4: Wishart density over $\mathrm{S}_{\mathcal{G}}$.

Example 11.2. Consider the DAG $\mathcal{G}$ given in Figure 4. Let us now apply the results of Theorem 11.2 to derive the density of $\Gamma \sim \pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$. First we need to find $\Sigma(\Gamma)^{\mathscr{V}}$, the completion of Γ in $\mathrm{PD}_{\mathcal{G}}$ restricted to the entries that correspond to $\mathcal{G}^m$. To this end, we only need to determine Σ_{32} (since all other sentries of Σ are already specified). The equation $\Sigma_{[3\succ} = \Sigma_{[3\succ}\Sigma^{-1}_{\prec 3\succ}\Sigma_{\prec 3\succ}$ and the fact that $pa(3) = \emptyset$ implies that $\Sigma_{32} = 0$. The next step is to compute $\mathrm{Iss}(\Sigma)$. Note that $\mathrm{Iss}(\Sigma)$ is a 6×6 matrix in $R^{\mathscr{W}\times\mathscr{W}}$, where $\mathscr{W} = \{(1,1),(2,1),(2,2),(3,1),(3,2),(3,3)\}$. Recall that

$$\mathrm{Iss}(\Sigma)_{ij,kl} = \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}, \quad (i,j),(k,l) \in \mathscr{W}.$$

and hence

$$\mathrm{Iss}(\Sigma) = \begin{pmatrix} 2\Sigma_{11}^2 & 2\Sigma_{21}\Sigma_{11} & 2\Sigma_{21}^2 & 2\Sigma_{31}\Sigma_{11} & 2\Sigma_{31}\Sigma_{21} & 2\Sigma_{31}^2 \\ 2\Sigma_{21}\Sigma_{11} & \Sigma_{22}\Sigma_{11} + \Sigma_{21}^2 & 2\Sigma_{22}\Sigma_{21} & \Sigma_{31}\Sigma_{21} & \Sigma_{31}\Sigma_{22} & 0 \\ 2\Sigma_{21}^2 & 2\Sigma_{22}\Sigma_{21} & 2\Sigma_{22}^2 & 0 & 0 & 0 \\ 2\Sigma_{31}\Sigma_{11} & \Sigma_{31}\Sigma_{21} & 0 & \Sigma_{33}\Sigma_{11} + \Sigma_{31}^2 & \Sigma_{33}\Sigma_{21} & 2\Sigma_{33}\Sigma_{31} \\ 2\Sigma_{31}\Sigma_{21} & \Sigma_{31}\Sigma_{22} & 0 & \Sigma_{33}\Sigma_{21} & \Sigma_{33}\Sigma_{22} & 0 \\ 2\Sigma_{31}^2 & 0 & 0 & 2\Sigma_{33}\Sigma_{31} & 0 & 2\Sigma_{33}^2 \end{pmatrix}.$$

Now for $E = \{(1,1),(2,1),(2,2),(3,1),(3,3)\}$ and $\mathscr{I} = \{(3,2)\}$, note that $\mathrm{Iss}(\Sigma)_{E|\mathscr{I}}$ is equal to

$$\begin{pmatrix} 2\Sigma_{11}^2 & 2\Sigma_{21}\Sigma_{11} & 2\Sigma_{21}^2 & 2\Sigma_{31}\Sigma_{11} & 2\Sigma_{31}^2 \\ 2\Sigma_{21}\Sigma_{11} & \Sigma_{22}\Sigma_{11} + \Sigma_{21}^2 & 2\Sigma_{22}\Sigma_{21} & \Sigma_{31}\Sigma_{21} & 0 \\ 2\Sigma_{21}^2 & 2\Sigma_{22}\Sigma_{21} & 2\Sigma_{22}^2 & 0 & 0 \\ 2\Sigma_{31}\Sigma_{11} & \Sigma_{31}\Sigma_{21} & 0 & \Sigma_{33}\Sigma_{11} + \Sigma_{31}^2 & 2\Sigma_{33}\Sigma_{31} \\ 2\Sigma_{31}^2 & 0 & 0 & 2\Sigma_{33}\Sigma_{31} & 2\Sigma_{33}^2 \end{pmatrix} - (\Sigma_{33}\Sigma_{22})^{-1} \begin{pmatrix} 2\Sigma_{31}\Sigma_{21} \\ \Sigma_{31}\Sigma_{22} \\ 0 \\ \Sigma_{33}\Sigma_{21} \\ 0 \end{pmatrix} \begin{pmatrix} 2\Sigma_{31}\Sigma_{21} \\ \Sigma_{31}\Sigma_{22} \\ 0 \\ \Sigma_{33}\Sigma_{21} \\ 0 \end{pmatrix}^t$$

From this we can compute $\det(\mathrm{Iss}(\Sigma)_{E|\mathscr{I}})$ explicitly. After some simplification the final expressions is given as follows:

$$\det(\mathrm{Iss}(\Sigma)_{E|\mathscr{I}}) = \frac{8\left(\Sigma_{33}\Sigma_{21}^2 + \Sigma_{22}\Sigma_{31}^2 - \Sigma_{11}\Sigma_{22}\Sigma_{33}\right)^4}{\Sigma_{22}\Sigma_{33}}.$$

The Jacobian calculation above allows us to specify the functional form of the density $\pi_{U,\alpha}^{\mathrm{S}_\mathcal{G}}(\Gamma)$ w.r.t to Lebesgue measure:

$$\begin{aligned} \pi_{U,\alpha}^{\mathrm{S}_\mathcal{G}}(\Gamma) &= z_\mathcal{G}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\} D_{11}^{-\frac{1}{2}\alpha_1+4} D_{22}^{-\frac{1}{2}\alpha_2+2} D_{33}^{-\frac{1}{2}\alpha_3+2} \\ &\times \left(\Sigma_{33}\Sigma_{21}^2 + \Sigma_{22}\Sigma_{31}^2 - \Sigma_{11}\Sigma_{22}\Sigma_{33}\right)^{-4} \Sigma_{22}\Sigma_{33} \\ &= z_\mathcal{G}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\} D_{11}^{-\frac{1}{2}\alpha_1+4} D_{22}^{-\frac{1}{2}\alpha_2+2} D_{33}^{-\frac{1}{2}\alpha_3+2} (\Sigma_{11|<1>})^{-4} (\Sigma_{22}\Sigma_{33})^{-4} \Sigma_{22}\Sigma_{33} \\ &= z_\mathcal{G}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\} D_{11}^{-\frac{1}{2}\alpha_1} D_{22}^{-\frac{1}{2}\alpha_2} D_{33}^{-\frac{1}{2}\alpha_3} \det(\Sigma_{<1>})^{-1}. \end{aligned} \tag{11.4}$$

The Isserlis matrix expressions in Proposition 11.2 provides a useful tool for computing the Jacobian of the mapping $(\Sigma^E \mapsto \Sigma^{-E}) : \mathrm{S}_\mathcal{G} \to \mathrm{R}_\mathcal{G}$. Nevertheless, Example 11.2 clearly demonstrates the complexity of the lengthy computations involved: even for the simplest of DAGs. A closer examination of Equation (11.4) suggests that the final expression for $\pi_{U,\alpha}^{\mathrm{S}_\mathcal{G}}(\Gamma)$ may be simplified in terms of the local properties of the DAG $\mathcal{G}$. To show that this is indeed the case we first prove the following lemma.

Lemma 11.1. *Let $\mathcal{G} = (V,E)$ be an arbitrary DAG, then the Jacobian of the mapping $(\Sigma^{-E} \mapsto \Sigma^E) : \mathrm{R}_\mathcal{G} \to \mathrm{S}_\mathcal{G}$ is given as follows:*

$$\prod_{i=1}^{m} \Sigma_{ii|<i>}^{pa_i+2} \det(\Sigma_{<i>}).$$

Proof. First note that the mapping $\Sigma^{-E} \mapsto \Sigma^E$ can be written as the composition of the two mappings $(\Sigma^{-E} \mapsto \times_{i=1}^m(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]}) : \mathrm{R}_{\mathcal{G}} \to \Xi_{\mathcal{G}}$ and $(\times_{i=1}^m(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]}) \mapsto \Sigma^E) : \Xi_{\mathcal{G}} \to \mathrm{S}_{\mathcal{G}}$. It is easy to check that the Jacobian of the first mapping is the same as the Jacobian of the inverse of the mapping $\psi : (L, D) \mapsto LD^{-1}L^t$ in Lemma 8.1 and is therefore equal to $\prod_{i=1}^m \Sigma_{ii|<i>}^{pa_i+2}$. So it remains to calculate the Jacobian of the second mapping.

We shall proceed by mathematical induction. Let us assume by the inductive hypothesis that the Jacobian of the mapping $(\times_{i=1}^m(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]}) \mapsto \Sigma^E) : \Xi_{\mathcal{G}} \to \mathrm{S}_{\mathcal{G}}$ is equal to $\prod_{i=1}^{|V|} \det(\Sigma_{<i>})$ for any DAG $\mathcal{G}$ with $|V| < m$. We will show that the result will also hold true for $|V| = m$. The case $m = 1$ is trivial. So assume that $m \geq 2$. Let $\mathcal{G}_{[1]}$ be the induced subgraph of $\mathcal{G}$ with the vertex set $V_{[1]} := V \setminus \{1\}$ and the corresponding edge set, denoted by $E_{[1]}$. Since $V_{[1]}$ is an ancestral subset of V, if Σ^E belongs to $\mathrm{S}_{\mathcal{G}}$, then $\Sigma^{E_{[1]}}$, the projection of Σ on $I_{G_{[1]}}$, is an element of $\mathrm{S}_{\mathcal{G}_{[1]}}$. Furthermore the positive definite completion of in $\mathrm{PD}_{\mathcal{G}_{[1]}}$ is indeed the principal sub-matrix $\Sigma_{V_{[1]}}$. The above two observations simply follow from the recursive nature of the completion process in Proposition 11.1). Now consider the following composition of the inverse mapping $\Sigma^E \mapsto \times_{i=1}^m(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]})$

$$\begin{array}{ccccc} \mathrm{S}_{\mathcal{G}} & \to & \mathbb{R}_+ \times \mathbb{R}^{<1]} \times \mathrm{S}_{\mathcal{G}_{[1]}} & \to & \mathbb{R}_+ \times \mathbb{R}^{<1]} \times \Xi_{\mathcal{G}_{[1]}} = \Xi_{\mathcal{G}} \\ \Sigma^E & \mapsto & (\Sigma_{11|<1>}, \Sigma_{<1>}^{-1}\Sigma_{<1]}, \Sigma^{E_{[1]}}) & \mapsto & \left(\Sigma_{11|<1>}, \Sigma_{<1>}^{-1}\Sigma_{<1]}, \times_{i=2}^m(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]})\right) \end{array}$$

By the inductive hypothesis the Jacobian of the second mapping,

$$(\Sigma_{11|<1>}, \Sigma_{<1>}^{-1}\Sigma_{<1]}, \Sigma^{E_{[1]}}) \mapsto \left(\Sigma_{11|<1>}, \Sigma_{<1>}^{-1}\Sigma_{<1]}, \times_{i=2}^m(\Sigma_{ii|<i>}, \Sigma_{<i>}^{-1}\Sigma_{<i]})\right),$$

is equal to $\prod_{i=2}^m \det(\Sigma_{<i>})^{-1}$. Hence it suffices to prove that the Jacobian of the first mapping,

$$\Sigma^E = (\Sigma_{11}, \Sigma_{<1]}, \Sigma^{E_{[1]}}) \mapsto (\Sigma_{11|<1>}, \Sigma_{<1>}^{-1}\Sigma_{<1]}, \Sigma^{E_{[1]}}) = (\Sigma_{11} - \Sigma_{[1>}\Sigma_{<1>}^{-1}\Sigma_{<1]}, \Sigma_{<1>}^{-1}\Sigma_{<1]}, \Sigma^{E_{[1]}})$$

is $\det(\Sigma_{<1>})^{-1}$. This follows by noting that the Jacobian matrix of this mapping is lower triangular and is given as follows:

$$\begin{pmatrix} I & 0 & 0 \\ * & \Sigma_{<1>}^{-1} & 0 \\ * & * & 1 \end{pmatrix}$$

Hence the results now follows by induction. □

We now proceed to state the functional form of the density of $\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{G}}}$ w.r.t Lebesgue measure without using Isserlis matrices.

Corollary 11.2. *Let $\mathcal{G} = (V, E)$ be an arbitrary DAG and let $\Sigma \sim \pi_{U,\alpha}^{\mathrm{PD}_{\mathcal{G}}}$. Now let $\Gamma = \mathrm{proj}(\Sigma) = \Sigma^E$, then the density of $\Gamma \sim \pi_{U,\alpha}^{\mathrm{S}_{\mathcal{G}}}$ w.r.t. Lebesgue measure is given by*

$$z_{\mathcal{G}}(U, \alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\} \prod_{i=1}^m \Sigma_{ii|<i>}^{-\frac{1}{2}\alpha_i} \det(\Sigma_{<i>})^{-1}. \tag{11.5}$$

Remark 11.4. In Remark 11.2 we established that when $\mathcal{G}$ is perfect $\mathrm{S}_{\mathcal{G}}$ and $\mathrm{Q}_{\mathcal{G}}$ are identical. Hence for $\mathcal{G}$ perfect $\pi_{U,\alpha}^{\mathrm{S}_{\mathcal{G}}}$ and $\pi_{U,\alpha}^{\mathrm{Q}_{\mathcal{G}}}$ are the same distribution.

We conclude this subsection by making the observation that the expression in Equation (11.5) is much simpler to evaluate than the expression in Equation (11.2).

11.3 The inverse DAG Wishart for homogeneous DAGs

In this paper we proceed to formally demonstrate that the class of inverse DAG Wisharts $\pi_{U,\alpha}^{S_{\mathcal{G}}}$ naturally contains an important sub-class of inverse Wishart distributions for that was introduced by Khare and Rajaratnam [9] in the context of Gaussian covariance graph models. In the process we also demonstrate that for a special class of DAGs the functional form of the density of the DAG Wisharts $\pi_{U,\alpha}^{S_{\mathcal{G}}}$ can be considerably simplified. Recall that a Gaussian covariance graph model over an undirected graph $G = (V, \mathscr{V})$, denoted by $\mathscr{N}(G_{\text{cov}})$, is defined as follows.

Definition 11.2. Let $\text{PD}_{G_{cov}}$ denote the set of positive definite matrices $\Sigma \in \text{PD}_m(\mathbb{R})$ such that $\Sigma_{ij} = 0$ whenever $i \not\sim_G j$, i.e., when i and j are not neighbors. Then the Gaussian covariance graph model over G is defined as

$$\mathscr{N}(G_{cov}) := \{\text{N}_m(0, \Sigma) : \Sigma \in \text{PD}_{G_{cov}}\}.$$

A formal comparison between the DAG Wishart priors introduced in this paper and the covariance Wishart priors introduced in [9] requires a few technical definitions.

Definition 11.3. *a*) A DAG $\mathcal{G}$ is called a homogeneous DAG of type I if it is transitive (i.e., $i \to j \to k$ implies that $i \to k$), and perfect. A DAG $\mathcal{G}$ is called a homogeneous DAG of type II if it is transitive and does not contain any induced subgraph of the form $j \leftarrow i \to k$.

b) An undirected graph $G = (V, \mathscr{V})$ is called homogeneous if for each pair of vertices $i, j \in V$,

$$i \sim_G j \Longrightarrow \text{ne}(i) \cup \{i\} \subseteq \text{ne}(j) \cup \{j\} \text{ or } \text{ne}(j) \cup \{j\} \subseteq \text{ne}(i) \cup \{i\}. \tag{11.6}$$

Equivalently, a graph G is said to be homogeneous if it is decomposable and does not contain the A_4 path as an induced subgraph. The reader is referred to [15] for further details on homogeneous graphs.

Note that if $\mathcal{G}$ is a homogeneous DAG of either types, then $\mathcal{G}^u$ is homogeneous. On the other hand, if $G = (V, \mathscr{V})$ is homogeneous, then one can construct a homogeneous DAG of type I or II that is a DAG version of G. This can be achieved by using the Hasse tree associated with the homogeneous (undirected) graph and using the given orientation to obtain a DAG of type I. Reversing the orientation (i.e., redirecting all the arrows to the root of the tree) will yield a DAG of type II. More precisely we shall now show an example that constructs a DAG version that is homogeneous of type II. Let $\mathcal{G}$ be a directed version of G obtained by directing each edge $i \sim_G j$ to a directed edge $i \to j$ if $\text{ne}(i) \cup \{i\} \subsetneq \text{ne}(j) \cup \{j\}$, or $j \to i$ if $\text{ne}(j) \cup \{j\} \subsetneq \text{ne}(i) \cup \{i\}$. If $\text{ne}(i) \cup \{i\} = \text{ne}(j) \cup \{j\}$, an arbitrary direction is chosen. From Equation (11.6) one can check that $\mathcal{G}$ is a transitive DAG and it does not contain any induced subgraph of the form $j \leftarrow i \to k$. In general, it can be shown that if $\mathcal{G}$ is a homogeneous DAG of type II and a DAG version of G, then $\mathscr{N}(\mathcal{G})$ is identical to the Gaussian covariance model $\mathscr{N}(G_{\text{cov}})$ in the sense that $\text{PD}_{G_{cov}} = \text{PD}_{\mathcal{G}}$ (see [18] for instance for more details). It is also evident, from the Markov equivalence of perfect DAGs and decomposable graphs, that for a homogeneous DAG $\mathcal{G}$ of type I which is a DAG version of G, we have $\text{PD}_G = \text{PD}_{\mathcal{G}}$.

Proposition 11.3. *Let $\mathcal{G} = (V, E)$ be a homogeneous DAG of either type I or II and let $G = (V, \mathscr{V})$ be a homogeneous graph.*

a) The density of $\pi_{U,\alpha}^{S_\mathcal{G}}$ is given by

$$z_\mathcal{G}(U,\alpha)^{-1} \exp\{-\frac{1}{2}\mathrm{tr}(\Sigma(\Gamma)^{-1}U)\} \prod_{i=1}^{m} \Sigma_{ii|<i>}^{-\frac{1}{2}(\alpha_i + 2ch_i(\mathcal{G}))}, \tag{11.7}$$

where $ch_i(\mathcal{G}) = |\mathrm{ch}_\mathcal{G}(i)|$.

b) If $\mathcal{G}$ is of type II and a DAG version of G, then the open cone $\mathrm{PD}_{G_{cov}}$ can be identified with $S_\mathcal{G}$ via the bijective mapping

$$\Gamma \mapsto [\Gamma]^0 := \Sigma(\Gamma) : S_\mathcal{G} \to \mathrm{PD}_{G_{cov}}. \tag{11.8}$$

Let $\pi_{U,\alpha}^{\mathrm{PD}_{G_{cov}}}$ denote the probability image of the inverse DAG Wishart $\pi_{U,\alpha}^{S_\mathcal{G}}$ under the mapping in Equation (11.8)*. Then the density of $\pi_{U,\alpha}^{\mathrm{PD}_{G_{cov}}}$ w.r.t. Lebesgue measure is given by Equation* (11.7)*.*

Proof. *a*) In light of Equation (11.5) in Corollary 11.2 it suffices to prove that for every $\Sigma \in \mathrm{PD}_\mathcal{G}$,

$$\prod_{i\in V} \det(\Sigma_{<i>}) = \prod_{i\in V} \Sigma_{ii|<i>}^{ch_i(\mathcal{G})}. \tag{11.9}$$

1. Suppose that $\mathcal{G}$ is homogeneous of type I. We shall first show that for every $i \in V$

$$\det(\Sigma_{<i>}) = \prod_{\ell\in \mathrm{pa}(i)} \Sigma_{\ell\ell|<\ell>}. \tag{11.10}$$

If $\mathrm{pa}(i) = \emptyset$ for some i, then by our convention $\det(\Sigma_{<i>}) = 1$ and $\Sigma_{\ell\ell|<\ell>} = 1$ for any $\ell \in \mathrm{pa}(i)$ and therefore Equation (11.10) holds. Now let ℓ_0 be the smallest integer in $\mathrm{pa}(i)$. One then can easily check that since $\mathcal{G}$ is both transitive and perfect we have $\mathrm{pa}(i) = \{\ell_0\} \cup \mathrm{pa}(\ell_0)$. From this we write

$$\det(\Sigma_{<i>}) = \Sigma_{\ell_0\ell_0|<\ell_0>} \det(\Sigma_{<\ell_0>}).$$

Now by repeating this procedure we obtain the result in Equation(11.10). Finally we write

$$\prod_{i\in V} \det(\Sigma_{<i>}) = \prod_{i\in V} \prod_{\ell\in \mathrm{pa}(i)} \Sigma_{\ell\ell|<\ell>} = \prod_{i\in V} \Sigma_{ii|<i>}^{ch_i(\mathcal{G})}.$$

2. Suppose $\mathcal{G}$ is homogeneous of type II. We shall proceed by induction. It is clear that Equation (11.9) holds when $m = |V| = 1$. Now by the inductive hypothesis assume that Equation (11.9) holds for every homogeneous DAG of type II, connected or disconnected, with fewer vertices than $m = |V|$. Using the inductive hypothesis we shall show that Equation (11.9) will also hold for $\mathcal{G}$ with m vertices. Now let $\Sigma \in \mathrm{PD}_\mathcal{G}$ be given.

Case 1) Suppose that $\mathcal{G}$ is connected. Let $\mathcal{D}$ be the induced DAG on $V \setminus \{1\}$. It is clear that $\mathcal{D}$ is a homogeneous DAG of type II and therefore by the induction hypothesis

$$\prod_{i=2}^{m} \det(\Psi_{<i>}) = \prod_{i=2}^{m} \Psi_{ii|<i>}^{ch_i(\mathcal{D})},$$

where $\Psi = \Sigma_{V\setminus\{1\}}$. Note that $\mathcal{D}$ is an ancestral subgraph of $\mathcal{G}$ and hence $\mathrm{fa}_{\mathcal{D}}(i) = \mathrm{fa}_{\mathcal{G}}(i)$ for each $i = 2, \ldots, m$ and consequently $\Psi_{<i>} = \Sigma_{<i>}$ and $\Psi_{ii|<i>} = \Sigma_{ii|<i>}$. All together these imply the following:

$$\prod_{i=2}^{m} \det(\Sigma_{<i>}) = \prod_{i=2}^{m} \Sigma_{ii|<i>}^{ch_i(\mathcal{D})}.$$

Now we claim that $\mathrm{fa}_{\mathcal{G}}(1) = V$. Assume to the contrary that $V \setminus \mathrm{fa}_{\mathcal{G}}(1) \neq \emptyset$. Since $\mathcal{G}$ is connected, this implies that there exist vertices $i \in \mathrm{fa}_{\mathcal{G}}(1)$ and $j \in V \setminus \mathrm{fa}_{\mathcal{G}}(1)$ such that i, j are adjacent in $\mathcal{G}$. But this implies $j \to i \to 1$ or $j \leftarrow i \to 1$. By definition these induced subgraphs cannot occur in $\mathcal{G}$. Thus $\preceq 1 \succeq = V$ and therefore we have

$$\det(\Sigma_{<1>}) = \Sigma_{11|<1>}^{-1} \det(\Sigma) = \prod_{i=2}^{m} \Sigma_{ii|<i>}.$$

Also the fact that $\mathrm{fa}_{\mathcal{G}}(1) = V$ implies that for each $i \in V \setminus \{1\}$ we have

$$ch_i(\mathcal{G}) = ch_i(\mathcal{D}) + 1.$$

Therefore

$$\begin{aligned}
\prod_{i\in V} \det(\Sigma_{<i>}) &= \det(\Sigma_{<1>}) \prod_{i=2}^{m} \det(\Sigma_{<i>}) \\
&= \prod_{i=2}^{m} \Sigma_{ii|<i>} \prod_{i=2}^{m} \Sigma_{ii|<i>}^{ch_i(\mathcal{D})} \\
&= \prod_{i\in V} \Sigma_{ii|<i>}^{ch_i(\mathcal{G})}.
\end{aligned}$$

Case 2) Suppose $\mathcal{G}$ is disconnected. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ denote respectively the induced subgraphs of $\mathcal{G}$ on $\mathrm{fa}_G(1)$ and $V \setminus \mathrm{fa}_{\mathcal{G}}(1)$. It is clear that $\mathcal{D}_1$ and $\mathcal{D}_2$ are both homogeneous of type II. In addition it is also easily verified that they are ancestral. Now let $\Psi := \Sigma_{\preceq 1 \succeq} \in \mathrm{PD}_{\mathcal{D}_1}$ and $\Psi' := \Sigma_{V\setminus \mathrm{fa}_{\mathcal{G}}(1)} \in \mathrm{PD}_{\mathcal{D}_2}$. Now applying the induction

hypothesis and the fact that $\mathcal{D}_1$ and $\mathcal{D}_2$ are disjoint we have:

$$\begin{aligned}
\prod_{i\in V}\det(\Sigma_{<i>}) &= \prod_{i\in \mathrm{fa}_{\mathcal{G}}(1)}\det(\Sigma_{<i>})\prod_{i\in V\setminus \mathrm{fa}_{\mathcal{G}}(1)}\det(\Sigma_{<i>}) \\
&= \prod_{i\in \mathrm{fa}_{\mathcal{G}}(1)}\det(\Psi_{<i>})\prod_{i\in V\setminus \mathrm{fa}_{\mathcal{G}}(1)}\det(\Psi'_{<i>}) \\
&= \prod_{i\in \mathrm{fa}_{\mathcal{G}}(1)}\det(\Psi_{ii|<i>})^{ch_i(\mathcal{D}_1)}\prod_{i\in V\setminus \mathrm{fa}_{\mathcal{G}}(1)}\det(\Psi'_{ii|<i>})^{ch_i(\mathcal{D}_2)} \\
&= \prod_{i\in \mathrm{fa}_{\mathcal{G}}(1)}\det(\Sigma_{ii|<i>})^{ch_i(\mathcal{G})}\prod_{i\in V\setminus \mathrm{fa}_{\mathcal{G}}(1)}\det(\Sigma'_{ii|<i>})^{ch_i(\mathcal{G})} \\
&= \prod_{i\in V}\det(\Sigma_{ii|<i>})^{ch_i(\mathcal{G})}.
\end{aligned}$$

b) It is clear that the mapping in Equation (11.8) is a diffeomorphism and the Jacobian of this mapping is 1. Thus the density $\pi^{\mathrm{PD}_{G_{cov}}}_{U,\alpha}$ w.r.t. Lebesgue measure is also given by equation (11.7). □

Remark 11.5. We note that for a homogeneous graph G the distribution $\pi^{\mathrm{PD}_{\mathcal{G}cov}}_{U,\alpha}$ with the associated density derived in Equation (11.7) coincides with the inverse Wishart distribution (or covariance Wishart priors) introduced by Khare and Rajaratnam [9].

11.4 Further properties of the DAG Wishart distributions $\pi^{\mathrm{R}_{\mathcal{G}}}_{U,\alpha}$ and $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$

We now proceed to derive useful properties of the DAG inverse Wishart distribution $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$. To this end, let us carefully lay out the setting. We begin with the Bayesian Gaussian model $\mathcal{N}(\mathcal{G})$ with parameter space $\mathrm{S}_{\mathcal{G}}$. Now the elements of $\mathcal{N}(\mathcal{G})$ are of the from $\mathrm{N}_m(0,\Sigma)$, such that $\Sigma^E \in \mathrm{S}_{\mathcal{G}}$. Therefore, if $\mathbf{x} \sim \mathrm{N}_m(0,\Sigma)$, then for each $i \in V$ the distribution of $\mathbf{x}_{ii|<i>}$ is parametrized by $(\Sigma_{ii|<i>}, \Sigma^{-1}_{<i>}\Sigma_{<i]})$. The following theorem formally establishes the strong directed hyper Markov property for the class $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$ for an arbitrary DAG.

Theorem 11.1. *Let $\mathcal{G} = (V,E)$ be an arbitrary DAG. If $\Sigma^E \sim \pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$, then*
i) *$\{(\Sigma_{ii|<i>}, \Sigma^{-1}_{<i>}\Sigma_{<i]} : i \in V\}$ are mutually independent and therefore $\pi^{\mathrm{S}_{\mathcal{G}}}_{U,\alpha}$ is strongly directed Markov.*
ii) *The distribution of $\Sigma_{ii|<i>}$ and $\Sigma^{-1}_{<i>}\Sigma_{<i]}|\Sigma_{ii|<i>}$ are, respectively, given by*

$$\Sigma_{ii|<i>} \sim IG(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1, \frac{1}{2}U_{ii|<i>}), \text{ and} \tag{11.11}$$

$$\Sigma^{-1}_{<i>}\Sigma_{<i]}|\Sigma_{ii|<i>} \sim \mathrm{N}_{pa_i}(U^{-1}_{<i>}U_{<i]}, \Sigma_{ii|<i>}U^{-1}_{<i>}). \tag{11.12}$$

Proof. The proof is omitted as it follows similarly to the one in Theorem 6.1. □

We now proceed to evaluate the expected value under $\pi^{\mathrm{S}_\mathcal{G}}_{U,\alpha}$. First note that since $\mathrm{S}_\mathcal{G}$ is an open subset of $\mathbb{R}^{|E|}$, the expected value of $\pi^{\mathrm{S}_\mathcal{G}}_{U,\alpha}$ is well defined.

Proposition 11.4. *Let $\mathcal{G}$ be an arbitrary DAG and $\Sigma^E \sim \pi^{\mathrm{S}_\mathcal{G}}_{U,\alpha}$, with $\alpha > pa_i + 4$. Then the expected value of Σ^E can be recursively computed by the following steps:*

(i) $\mathbb{E}[\Sigma_{mm}] = \dfrac{U_{mm}}{\alpha_m - 4}$,

(ii) $\mathbb{E}[\Sigma_{<i]}] = -\mathbb{E}[\Sigma_{<i>}]U^{-1}_{<i>}U_{<i]}$,

(iii) $\mathbb{E}(\Sigma_{ii}) = \dfrac{U_{ii|<i>}}{\alpha_i - pa_i - 4} + \mathrm{tr}\left(\mathbb{E}[\Sigma_{<i>}]\left(\dfrac{U_{ii|<i>}U^{-1}_{<i>}}{\alpha_i - pa_i - 4} + U^{-1}_{<i>}U_{<i]}U_{[i>}U^{-1}_{<i>}\right)\right)$, $i = m-1, \ldots, 1$.

Proof. Since Equation (11.11) and Equation (11.12) are analogous versions of Equation (6.1) and Equation (6.2), but for general DAGs, the proof follows along the same lines as the proof in Proposition 8.2, and is therefore omitted. □

We now proceed to analyze the DAG Wishart distribution $\pi^{\mathrm{R}_\mathcal{G}}_{U,\alpha}$ as a class of distributions in their own right. Once more let $\mathcal{G}$ be an arbitrary DAG and α a given vector in $\mathbb{R}^m$ such that $\alpha_i > pa_i + 2$, $\forall i$. Now consider the family of DAG Wishart disitrbutions $\{\pi^{\mathrm{R}_\mathcal{G}}_{U,\alpha} : U \in \mathrm{PD}_\mathcal{G}\}$. Since $\mathrm{PD}_\mathcal{G}$ is isomorphic to $\mathcal{S}_\mathcal{G}$ via the mapping $U \mapsto U^E$, it is more natural to parametrize this family of distributions as $\{\pi^{\mathrm{R}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$. It is easy to check that this is an identifiable parametrization, i.e., if $\pi^{\mathrm{R}_\mathcal{G}}_{U^E_1,\alpha}$ is a.s. equal to $\pi^{\mathrm{R}_\mathcal{G}}_{U^E_1,\alpha}$, then $U^E_1 = U^E_2$. The following lemma formalizes these points.

Lemma 11.2. *Let $\mathcal{G}$ be a perfect DAG and let α be given. Then the Wishart family $\{\pi^{\mathrm{R}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$, or equivalently $\{\pi^{\mathrm{P}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$, is a general exponential family. If $\mathcal{G}$ is not a perfect DAG then $\{\pi^{\mathrm{R}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$ is no longer a general exponential family but a curved exponential family.*

Proof. Let $t : \mathrm{R}_\mathcal{G} \to \mathrm{Z}_\mathcal{G}$ be the embedding $\Upsilon \mapsto [\Upsilon]^0$ and let $\eta : \mathrm{S}_\mathcal{G} \to \mathrm{Z}_\mathcal{G}$ be the embedding $U^E \mapsto \left[U^E\right]^0$. Then $\mathrm{tr}(\widehat{\Upsilon}U)$ is equal to the inner product of $[\Upsilon]^0$ and $\left[U^E\right]^0$ in Euclidian space $\mathrm{Z}_\mathcal{G}$. Note also that under these natural embeddings both $\mathrm{R}_\mathcal{G}$ and $\mathrm{S}_\mathcal{G}$ are open subsets of $\mathrm{Z}_\mathcal{G}$. The result that $\{\pi^{\mathrm{P}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$, is a general exponential family follows immediately from these observations.

Now if $\mathcal{G}$ is not perfect, the expression $\mathrm{tr}(\widehat{\Upsilon}U)$ not only depends on the entries in position ij where i, j are adjacent in $\mathcal{G}$, but also on a position ij where there exists an immorality $i \to k \leftarrow j$. Therefore, $\mathrm{tr}(\widehat{\Upsilon}U)$ is not equal to $\mathrm{tr}([\Upsilon]^0\left[U^E\right]^0)$, the inner product of $[\Upsilon]^0$ and $\left[U^E\right]^0$ in $\mathrm{Z}_\mathcal{G}$. It is however is clear that $\mathrm{tr}(\widehat{\Upsilon}U)$ is the inner product of the projection of $\widehat{\Upsilon}$ and U in Euclidean space $\mathrm{Z}_{\mathcal{G}^\mathrm{m}}$, which has higher dimension than $|E|$. Hence when $\mathcal{G}$ is not perfect $\{\pi^{\mathrm{R}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$ is no longer an exponential family, but only a curved exponential family.

□

Note that the proof of Lemma 11.2 shows that for an arbitrary non-perfect DAG $\mathcal{G}$, the family of DAG Wishart distributions $\{\pi^{\mathrm{R}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$ is strictly a subfamily of $\{\pi^{\mathrm{R}_\mathcal{G}}_{U,\alpha} :$

$U \in \mathrm{PD}_m(\mathbb{R})\}$. On the other hand, if $\mathcal{G}$ is perfect, then $\{\pi^{\mathrm{R}_\mathcal{G}}_{U,\alpha} : U \in \mathrm{PD}_m(\mathbb{R})\}$ is identical to $\{\pi^{\mathrm{R}_\mathcal{G}}_{U^E,\alpha} : U^E \in \mathrm{S}_\mathcal{G}\}$.

12 Closing remarks

This paper introduces a class of multi-parameter hyper Markov laws which generalize the classical Wishart distribution in a way that is useful for Bayesian inference for Gaussian directed acyclic graph (DAG) models. The paper then proceeds to develop a theoretical framework for Bayesian inference for DAG models in the Gaussian setting. The main breakthrough that has been achieved is that the framework applies to all DAG models and not just the narrower class of perfect DAGs. The perfect or decomposable assumption, a common feature in theoretical analysis of concentration and covariance graph models, tends to yield more abstract results, as compared to practical procedures. The development undertaken in this paper is free of such assumptions as it applies to all DAG models. This of course has tremendous benefits for applications in high dimensional settings. More specifically, the class of DAG Wishart distributions that are developed and investigated in this paper yields a rich and flexible class of conjugate Wishart distributions which generalize previous Wishart type distributions introduced in the literature. We proceed to demonstrate that normalizing constants, hyper-Markov properties, moments and Laplace transforms are available in closed form for our family of DAG Wisharts. Sampling from the distribution also does not resort to expensive computational techniques - resulting in inferential procedures that are scalable to very high dimensional problems.

Despite the advantages of this class of DAG Wishart distributions, we demonstrate that it is a challenge to evaluate their densities on the space of covariance and concentration spaces, as these are curved manifolds. In particular, covariance and concentration spaces for non-perfect DAGs correspond to non-Euclidean spaces, on which densities w.r.t standard Lebesgue measure are not defined. The results in this paper develops two approaches to deriving priors on covariance and concentration spaces corresponding to arbitrary non-perfect DAGs. In the process classes of DAG Wishart and DAG inverse Wishart distributions have been introduced and studied. Moreover, posterior moments are derived and shown to be in closed form. The theory that is developed is also illustrated through examples to demonstrate that the methodology is readily applicable.

References

[1] Andersson, S., Perlman, M. (1998). Normal linear regression models with recursive graphical Markov structure. *J. Multivariate Anal.*, **66**, 133–187.

[2] Billingsley, Patrick. (1979). *Probability and measure*, 2*nd* ed. Wiley.

[3] Ben-David, E., Rajaratnam, B.(2010). Generalized Hyper Markov laws for directed acyclic graphs-II ,*Technical Report*, Department of Statistics, Stanford University, *submitted*.

[4] Ben-David, E., Rajaratnam, B.(2010). On the Latec & Massam's conjecture for decomposable graph models, *in press*, Department of Statistics, Stanford University.

[5] Ben-David, E., Rajaratnam, B., (2010). Positive definite completion for Acyclic Digraphs, *Technical Report*, Department of Statistics, Stanford University.

[6] Cowell, G.R., Dawid, A.P., Lauritzen, L.S. and Spigelhalter, D. (1999). *Probabilistic networks and expert systems*, Springer-Verlag New York, Inc.

[7] Dawid, A.P. and Lauritzen, S.L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models, *Ann. Statist.* **21**, 1272-1317.

[8] Grőne, R., Johnson, C.R., Sa, E.M. and Wolkowicz, H. (1984). Positive definite completions of partial hermitian matricies, *Linear Algebra Appl.* **58**, 109–124.

[9] Khare, K. and Rajaratnam, B. (2010). Wishart distributions for decomposable covariance graph models, *Ann. Statist*, *in press*.

[10] Khare, K. and Rajaratnam, B. (2009). Bayesian covariance estimation in covariance graph models, *in press*, Department of Statistics, Stanford University.

[11] Lauritzen, S. L., Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their applications to expert systems, *Statist. Soc. Ser. B* **50**, pp. 157–224.

[12] Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H. G. (1990). Independence properties of directed Markov fields, *Networks.* **20**, pp. 491–505.

[13] Lauritzen, S.L. (1996). *Graphical models*, Oxford University Press Inc., New York.

[14] Lauritzen, S. L. and Spiegelhalter, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion), *J. R. Statist. Soc.* Ser. B **50**, pp. 157–224.

[15] Letac, G. and Massam, H. (2007). Wishart distributions for decomposable graphs, *Ann. Statist.* **35**, pp. 1278–1323.

[16] Muirhead, R.J. (1982). *Aspects of multivariate statistical theory*, John Wiley and Sons, New York.

[17] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann, San Mateo, CA.

[18] Pearl J. amd Wermuth N. (1993). When can association graphs admit a causal interpretation? *I*n Selecting Models from Data: Artificial Intelligence and Statistics IV, volume 89 of Lecture Notes in Statistics, pp 205214. Springer, New York.

[19] Rajaratnam, B., Massam, H. and Carvalho, C. (2008). Flexible covariance estimation in graphical models, *Ann. Statist.* **36**, pp. 2818-2849.

[20] Roverato, A. and Whitakar, J. (1998). The Isserlis matrix and its application to non-decomposable graphical Gaussian models , *Biometrika* **85**, pp. 711–725.

[21] Roverato, A. (2000). Cholesky decomposition of an inverse Wishart matrix, *Biometrika* **87**, pp. 99–112.

[22] Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion), *Statist. Sci.* **8**, pp. 219–283.

[23] Temme, N. M. (1996).
textitSpecial functions: An Introduction to the Classical Functions of Mathematical Physics, Wiley, New York.

[24] Verma, T. and Pearl, J., Causal networks: Semantics and expressiveness, in R. D. Shachter, T. S. Levitt, L. N. Kanal and J.F. Lemmer (eds), *Uncertainty in Artificial Intellegence 4*, North-Holland, Amsterdam, The Netherlands, pp.69–76.

[25] Wermuth, N. (1980). Linear recursive equations, covariance selection, and path analysis, *J. Amer. Statist. Assoc.* **75**, pp. 963–972.

Appendix/Supplemental Section

Proof of Theorem 5.1. Let us first simplify the expression by integrating out the terms involving D_{ii}'s.

$$
\begin{aligned}
& \int \exp\{-\frac{1}{2}\mathrm{tr}\left(\left(LD^{-1}L^t\right)U\right)\}\prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i} dLdD \\
= & \int \exp\{-\frac{1}{2}\mathrm{tr}\left(D^{-1}\left(L^tUL\right)\right)\}\prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i} dLdD \\
= & \int \exp\{-\frac{1}{2}\sum_{i=1}^{m} D_{ii}^{-1}(L^tUL)_{ii}\}\prod_{i=1}^{m} D_{ii}^{-\frac{1}{2}\alpha_i} dDdL \\
= & \int\left(\prod_{i=1}^{m}\int \exp\{-\frac{1}{2}D_{ii}^{-1}(L^tUL)_{ii}\}D_{ii}^{-\frac{1}{2}\alpha_i} dD_{ii}\right)dL \\
= & \int\prod_{i=1}^{m}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{((L^tUL)_{ii})^{\frac{\alpha_i}{2}-1}}dL \qquad \text{(if and only if } \alpha_i > 2\ \forall\ i = 1,2,\cdots,m) \\
= & \int\prod_{i=1}^{m}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{((L_{\cdot i})^tUL_{\cdot i})^{\frac{\alpha_i}{2}-1}}dL \\
= & \int\prod_{i=1}^{m}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{\left(\begin{pmatrix}1 & L_{<i]}^t\end{pmatrix}\begin{pmatrix}U_{ii} & U_{[i>}\\ U_{<i]} & U_{<i>}\end{pmatrix}\begin{pmatrix}1\\ L_{<i]}\end{pmatrix}\right)^{\frac{\alpha_i}{2}-1}}dL \\
= & \prod_{i=1}^{m}\int_{\mathbb{R}^{pa_i}}\frac{\Gamma\left(\frac{\alpha_i}{2}-1\right)2^{\frac{\alpha_i}{2}-1}}{\left(\begin{pmatrix}1 & L_{<i]}^t\end{pmatrix}\begin{pmatrix}U_{ii} & U_{[i>}\\ U_{<i]} & U_{<i>}\end{pmatrix}\begin{pmatrix}1\\ L_{<i]}\end{pmatrix}\right)^{\frac{\alpha_i}{2}-1}}dL_{<i]}. \qquad \text{eqn(A)}
\end{aligned}
$$

We now show how in general one can evaluate an integral of the form

$$\int_{\mathbb{R}^d}\frac{d\mathbf{x}}{\left(\begin{pmatrix}1 & \mathbf{x}^t\end{pmatrix}\begin{pmatrix}a & \mathbf{b}^t\\ \mathbf{b} & A\end{pmatrix}\begin{pmatrix}1\\ \mathbf{x}\end{pmatrix}\right)^{\gamma}},$$

where the block partitioned matrices, formed by $a \in \mathbb{R}$, $\mathbf{b} \in \mathbb{R}^d$ and the $(d-1)\times(d-1)$ matrix A, is positive definite. In order to simplify the above integral we proceed in two steps.

1) We first note that by the formula provided on [**?**, page 16] that,

$$\int_{\mathbb{R}}\frac{1}{(1+x^2)^{\gamma}}dx = \begin{cases}\frac{\sqrt{\pi}\Gamma\left(\gamma-\frac{1}{2}\right)}{\Gamma(\gamma)} & \gamma > \frac{1}{2},\\ \infty & \text{otherwise.}\end{cases}$$

By repeated application, we can generalize the above formula to

$$\int_{\mathbb{R}^d} \frac{1}{(\mathbf{x}^t\mathbf{x}+1)^\gamma} d\mathbf{x} = \begin{cases} \frac{(\sqrt{\pi})^d \Gamma\left(\gamma - \frac{d}{2}\right)}{\Gamma(\gamma)} & \gamma > \frac{d}{2}, \\ \infty & \text{otherwise.} \end{cases}$$

2) Let us now consider the general integral

$$\int_{\mathbb{R}^d} \frac{d\mathbf{x}}{\left(\begin{pmatrix}1 & \mathbf{x}^t\end{pmatrix}\begin{pmatrix}a & \mathbf{b}^t \\ \mathbf{b} & A\end{pmatrix}\begin{pmatrix}1 \\ \mathbf{x}\end{pmatrix}\right)^\gamma}.$$

Making the linear transformation $\mathbf{y} = A^{\frac{1}{2}}\mathbf{x} + A^{-\frac{1}{2}}\mathbf{b}$ it follows that for $\gamma > \frac{d}{2}$,

$$\begin{aligned} \int_{\mathbb{R}^d} \frac{d\mathbf{x}}{\left(\begin{pmatrix}1 & \mathbf{x}^t\end{pmatrix}\begin{pmatrix}a & \mathbf{b}^t \\ \mathbf{b} & A\end{pmatrix}\begin{pmatrix}1 \\ \mathbf{x}\end{pmatrix}\right)^\gamma} &= \frac{1}{\det(A)^{\frac{1}{2}}} \int_{\mathbb{R}^d} \frac{1}{(\mathbf{y}^t\mathbf{y} + a - \mathbf{b}^t A^{-1}\mathbf{b})^\gamma} d\mathbf{y} \\ &= \frac{(\sqrt{\pi})^d \Gamma\left(\gamma - \frac{d}{2}\right)}{\Gamma(\gamma)\det(A)^{\frac{1}{2}}\,(a - \mathbf{b}^t A^{-1}\mathbf{b})^{\gamma - \frac{d}{2}}}. \end{aligned} \tag{12.1}$$

Applying the result from Equation (12.1) to the desired integral in Equation(A) we obtain

$$\begin{aligned} z_{\mathcal{G}}(U,\alpha) &= \prod_{i=1}^m \int_{\mathbb{R}^{pa_i}} \frac{\Gamma\left(\frac{\alpha_i}{2} - 1\right) 2^{\frac{\alpha_i}{2}-1}}{\left(\begin{pmatrix}1 & L_{<i]}^t\end{pmatrix}\begin{pmatrix}U_{ii}U_{[i>} & \\ U_{<i]} & U_{<i>}\end{pmatrix}\begin{pmatrix}1 \\ L_{<i]}\end{pmatrix}\right)^{\frac{\alpha_i}{2}-1}} dL_{<i]} \\ &= \prod_{i=1}^m \frac{\Gamma\left(\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1\right) 2^{\frac{\alpha_i}{2}-1} (\sqrt{\pi})^{pa_i} \det(U_{<i>})^{\frac{\alpha_i}{2} - \frac{pa_i}{2} - \frac{3}{2}}}{\det(U_{\leq i \geq})^{\frac{\alpha_i}{2} - \frac{pa_i}{2} - 1}}, \end{aligned}$$

where $\det(U_{<i>}) := 1$ whenever $pa(i) = \emptyset$. It is easily seen that $z_{\mathcal{G}}(U,\alpha)$ is finite if and only if $\alpha_i > pa_i + 2$ for each $i = 1, \ldots, m$. □

Proof of Lemma 5.1. The likelihood of the data is given by

$$f(\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n \mid L, D) = \frac{1}{(\sqrt{2\pi})^{nm}} \exp\{-\frac{1}{2}\mathrm{tr}\left(LD^{-1}L^t(nS)\right)\} \det(D)^{-\frac{1}{2}n}.$$

When using $\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}$ as the prior for (D, L), the posterior distribution of (D, L) given the data $(\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n)$ is given by

$$\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}(L, D \mid \mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n) \propto \exp\{-\frac{1}{2}\mathrm{tr}\left(LD^{-1}L^t(nS + U)\right)\} \prod_{i=1}^m D_{ii}^{-\frac{n+\alpha_i}{2}}, \quad (D, L) \in \Theta_{\mathcal{G}}. \tag{12.2}$$

Hence the functional form of the posterior density is the same as that of the prior density, i.e.,

$$\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}(\cdot \mid \mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_n) = \pi_{\widetilde{U},\widetilde{\alpha}}(\cdot),$$

where $\widetilde{U} = nS + U$ and $\widetilde{\alpha} = (\alpha_1 + n, \ldots, \alpha_m + n)$.

Proof of Theorem 6.1. First consider the bijective mapping from the Cholesky parameterization to the D-parameterization:

$$\phi := ((D, L) \mapsto \times_{i \in V}(D_{ii}, L_{<i]})) : \Theta_{\mathcal{G}} \to \Xi_{\mathcal{G}}, \tag{12.3}$$

with the inverse mapping $(\times_{i \in V}(\lambda_i, \beta_{<i]}) \mapsto (D, L)) : \Xi_{\mathcal{G}} \to \Theta_{\mathcal{G}}$, where $D = \mathrm{diag}(\lambda_1 \ldots, \lambda_m)$ and

$$L_{ij} = \begin{cases} 1 & i = j \\ L_{ij} = \beta_{ij} & i \in pa(j) \\ 0 & \text{otherwise} \end{cases}$$

Note that $\beta_{<j]} = (\beta_{ij} : i \in pa(j))$ belongs to $\mathbb{R}^{<j]}$. Now $\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}$ naturally induces a prior on $\Xi_{\mathcal{G}}$ which we shall denote by $\pi_{U,\alpha}^{\Xi_{\mathcal{G}}}$. As noted in Remark 4.2 ϕ in Equation (12.3) is simply a permutation of the entries of D and L, hence its Jacobian is equal to 1. To derive the density of $\pi_{U,\alpha}^{\Xi_{\mathcal{G}}}$ it suffices to find an expression for $\mathrm{tr}((LD^{-1}L^t)U)$ in terms of $\prod_{i \in V}(D_{ii}, L_{<i]})$. To this end, we proceed as follows.

$$\begin{aligned}
&\mathrm{tr}((LD^{-1}L^t)U) = \mathrm{tr}(D^{-1}L^t)UL) = \sum_{i \in V} D_{ii}^{-1}(L^tUL)_{ii} \\
&= \sum_{i \in V} D_{ii}^{-1}(\sum_{k,l \in V} L_{ki}U_{kl}L_{li}) \\
&= \sum_{i \in V} D_{ii}^{-1} \begin{pmatrix} 1 \\ L_{<i]} \end{pmatrix}^t \begin{pmatrix} U_{ii} & U_{[i>} \\ U_{<i]} & U_{<i>} \end{pmatrix} \begin{pmatrix} 1 \\ L_{<i]} \end{pmatrix} \\
&= \sum_{i \in V} D_{ii}^{-1}(U_{ii} + L_{<i]}^tU_{<i]} + U_{[i>}L_{<i]} + L_{<i]}^tU_{<i>}L_{<i]}) \\
&= \sum_{i \in V} \left(D_{ii}^{-1}(L_{<i]} + U_{<i>}^{-1}U_{<i]})^tU_{<i>}(L_{<i]} + U_{<i>}^{-1}U_{<i]}) + D_{ii}^{-1}U_{ii|<i>}\right).
\end{aligned}$$

Therefore, the density of $\pi_{U,\alpha}^{\Xi_{\mathcal{G}}}$ w.r.t. the Lebesgue measure $\prod_{i \in V} d\lambda_i d\beta_{<i]}$ on $\times_{i \in V}(\mathbb{R}_+, \mathbb{R}^{<i]})$ is given by

$$z_{\mathcal{G}}(\alpha, U)^{-1} \exp\{-\frac{1}{2}\sum_{i \in V}\left(\lambda_i^{-1}(\beta_{<i]} + U_{<i>}^{-1}U_{<i]})^tU_{<i>}(\beta_{<i]} + U_{<i>}^{-1}U_{<i]}) + \lambda_i^{-1}U_{ii|<i>}\right)\} \prod_{i \in V} \lambda_i^{-\frac{1}{2}\alpha_i}. \tag{12.4}$$

The above clearly shows that $\{(\lambda_i, \beta_{<i]}) : i = 1, \ldots, m\}$ are mutually independent. To complete the proof we first integrate out $\beta_{<i]}$ to obtain the marginal density of λ_i. Notice that the expression involving $\beta_{<i]}$ in Equation (12.4) is an unnormalized multivariate normal integral and thus Equation (12.4) can be expressed as follows:

$$\int_{\mathbb{R}^{<i]}} \exp\{-\frac{1}{2}\sum_{i \in V} \lambda_i^{-1}(\beta_{<i]} + U_{<i>}^{-1}U_{<i]})^tU_{<i>}(\beta_{<i]} - U_{<i>}^{-1}U_{<i]}) + \lambda_i^{-1}U_{ii|<i>}\} \prod_{i \in V} \lambda_i^{-\frac{1}{2}\alpha_i} d(\beta_{<i]}) \tag{12.5}$$

$$\propto \exp\{-\frac{1}{2}\lambda_i^{-1}U_{ii|<i>}\} \prod_{i \in V} \lambda_i^{-\frac{1}{2}\alpha_i + \frac{1}{2}pa_i}$$

The above shows that $\lambda_i \sim IG(\alpha_i/2 - pa_i/2 - 1, U_{ii|<i>}/2)$. It is evident from Equation (12.5) that $\beta_{<i]}|\lambda_i \sim \mathrm{N}_{pa_i}(-U_{<i>}^{-1}U_{<i]}, \lambda_i U_{<i>}^{-1})$. It is also immediately clear that the same result holds for elements of the Cholesky-parameterization $(D_{ii}, L_{<i]})$, $i = 1, 2, \ldots, m$ as specified in the statement of the theorem.

Proof of Corollary 6.1. From Theorem 6.1:

$$
\begin{aligned}
&\int \frac{1}{(2\pi)^{pa_i/2}\det(D_{ii}U_{<i>}^{-1})^{1/2}} \exp\{(L_{<i]} + U_{<i>}^{-1}U_{<i]})^t (D_{ii}^{-1}U_{<i>})(L_{<i]} + U_{<i>}^{-1}U_{<i]})\} \\
&\times \frac{(1/2U_{ii|<i>})^{\alpha_i/2-pa_i/2-1}}{\Gamma(\alpha_i/2 - pa_i/2 - 1)} D_{ii}^{-\alpha_i/2+pa_i/2} \exp\{-1/2U_{ii|<i>}D_{ii}^{-1}\} dD_{ii} \\
=&\frac{\det(U_{<i>})^{1/2}(U_{ii|<i>})^{\alpha_i/2-pa_i/2-1}}{2^{\alpha_i/2-1}\pi^{pa_i/2}\Gamma(\alpha_i/2 - pa_i/2 - 1)} \int D_{ii}^{-\alpha_i/2} \exp\{-u_i D_{ii}^{-1}\} dD_{ii} \\
=&\frac{\det(U_{<i>})^{1/2}(U_{ii|<i>})^{\alpha_i/2-pa_i/2-1}}{2^{\alpha_i/2-1}\pi^{pa_i/2}\Gamma(\alpha_i/2 - pa_i/2 - 1)} \times \frac{\Gamma(\alpha_i/2 - 1)}{u_i^{\alpha_i/2-1}},
\end{aligned}
$$

where $u_i = 1/2U_{ii|<i>} + (L_{<i]} + U_{<i>}^{-1}U_{<i]})^t U_{<i>}(L_{<i]} + U_{<i>}^{-1}U_{<i]})$. Therefore the density of $L_{<i]}$ is given by

$$
c_i \left[1/2U_{ii|<i>} + (L_{<i]} + U_{<i>}^{-1}U_{<i]})^t U_{<i>}(L_{<i]} + U_{<i>}^{-1}U_{<i]})\right]^{-\alpha_i/2+1}. \tag{12.6}
$$

By Theorem 6.1 $L_{<i]}$ are mutually independent, hence the form of the density in the statement of the corollary is immediate from the above calculations. The parameters corresponding to the t-distribution follow by comparing the density in Equation (12.6) to the functional form of the density of the multivariate t-distribution.

Proof of Corollary 7.1.
By definition, the Laplace transform of $(\lambda, \mathbf{x})$ at $(\xi, u) \in \mathbb{R} \times \mathbb{R}^p$ is

$$
\begin{aligned}
&\int \exp\{-(\lambda\xi + u^t x)\} d\mathrm{N}_p(\mu, \lambda\Psi)(x) d\mathrm{IG}(\nu, \eta)(\lambda) \\
=&\int \exp\{-\lambda\xi\} \left(\int \exp\{-u^t x)\} d\mathrm{N}_p(\mu, \lambda\Psi)(x) \right) d\mathrm{IG}(\nu, \eta)(\lambda) \\
=&\int \exp\{-\lambda\xi\} \exp\{-u^t\mu + \frac{1}{2}\lambda u^t\Psi u\} d\mathrm{IG}(\nu, \eta)(\lambda) \\
=&\int \exp\{-\lambda\xi\} \exp\{-u^t\mu + \frac{1}{2}\lambda u^t\Psi u\} \left(\frac{\eta^\nu}{\Gamma(\nu)} \exp\{-\eta\lambda^{-1}\}\lambda^{-\nu-1} \right) d\lambda \\
=&\frac{\eta^\nu}{\Gamma(\nu)} \exp\{-u^t\mu\} \int \exp\{-(\xi - \frac{1}{2}u^t\Psi u)\lambda - \eta\lambda^{-1}\}\lambda^{-\nu-1} d\lambda \\
=&\frac{2\eta^\nu}{\Gamma(\nu)} \exp\{-u^t\mu\} \left(\frac{\xi - \frac{1}{2}u^t\Psi u}{\eta} \right)^{\frac{1}{2}\nu} K_\nu \left(2\sqrt{\eta(\xi - \frac{1}{2}u^t\Psi u)} \right) \\
=&\frac{2}{\Gamma(\nu)} \exp\{-u^t\mu\} \left(\eta(\xi - \frac{1}{2}u^t\Psi u) \right)^{\frac{1}{2}\nu} K_\nu \left(2\sqrt{\eta(\xi - \frac{1}{2}u^t\Psi u)} \right).
\end{aligned}
$$

Note that in computing the integral above we have used the fact that the Laplace transform of $\mathrm{N}_p(\mu, \lambda\Psi)$ at u is equal to $\exp\{-u^t\mu + \frac{1}{2}\lambda u^t\Psi u\}$. For computing the integral w.r.t. $d\lambda$ we use the Equation (9.42) in [23, page 235].

Proof of Lemma 7.1.

By definition, the Laplace transform of $\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}$ at $(\Lambda, Z) \in \Theta_{\mathcal{G}}$ is given by

$$\mathcal{L}_{\Theta_{\mathcal{G}}}(\Lambda, Z) := \int \exp\{-\mathrm{tr}(\Lambda D^t) - \mathrm{tr}(ZL^t)\}\pi_{U,\alpha}^{\Theta_{\mathcal{G}}}(D, L)dDdL.$$

Now under the change of variable $\phi : \Theta_{\mathcal{G}} :\to \Xi_{\mathcal{G}}$ defined in Equation (12.3) and the fact that

$$\mathrm{tr}(\Lambda D^t) + \mathrm{tr}(ZL^t) = \sum_i^m D_{ii}\Lambda_{ii} + \sum_{i=1}^m \left(1 + L_{<i]}^t Z_{<i]}\right)$$

we have

$$\begin{aligned}\mathcal{L}_{\Theta_{\mathcal{G}}}(\Lambda, Z) &= \int \exp\{-\sum_i^m D_{ii}\Lambda_{ii} - \sum_{i=1}^m \left(1 + L_{<i]}^t Z_{<i]}\right)\}\pi_{U,\alpha}^{\Xi_{\mathcal{G}}}(\times_{i=1}^m (D_{ii}, L_{<i]}))\prod_{i=1}^m dD_{ii}dL_{<i]} \\ &= e^{-m}\mathcal{L}_{\Xi_{\mathcal{G}}}(\times_{i=1}^n(\Lambda_{ii}, Z_{<i]})).\end{aligned}$$

Proof of Lemma 8.1.

Let $\Omega \in \mathrm{P}_{\mathcal{G}}$, and $(D, L) \in \Theta_{\mathcal{G}}$ such that $\Omega = LD^{-1}L^t$. Note that for each $1 \le j \le i \le m$,

$$\Omega_{ij} = (LD^{-1}L^t)_{ij} = \sum_{k=1}^m L_{ik}L_{jk}D_{kk}^{-1} = \sum_{k=1}^j L_{ik}L_{jk}D_{kk}^{-1}, \tag{12.7}$$

since L is lower triangular. Now from Equation (12.7) it follows by noting that $L_{jj} = 1, \forall\, j$,

$$\frac{\partial}{\partial L_{ij}}(LD^{-1}L^t)_{ij} = D_{jj}^{-1}, \;\; (i, j) \in E, \qquad \frac{\partial}{\partial D_{ii}}(LD^{-1}L^t)_{ii} = -D_{ii}^{-2}, \;\; i = 1, 2, \cdots, m.$$

Arrange the entries of $\Theta = (D, L) \in \Theta_{\mathcal{G}}$ as $D_{11}, \{L_{2k} : \; (2, k) \in E, 1 \le k < 2\}, D_{22}, \{L_{3k} : (3, k) \in E, 1 \le k < 3\}, \cdots, D_{m-1,m-1}, \{L_{mk} : \; (m, k) \in E, 1 \le k < m\}, D_{mm}$, and the entries of $\Omega \in \mathrm{P}_{\mathcal{G}}$ as $\Omega_{11}, \{\Omega_{2k} : \; (2, k) \in E, 1 \le k < 2\}, \Omega_{22}, \{\Omega_{3k} : \; (3, k) \in E, 1 \le k < 3\}, \cdots, \Omega_{m-1,m-1}, \{\Omega_{mk} : \; (m, k) \in E, 1 \le k < m\}, \Omega_{mm}$. From (12.7) it is easily seen that Ω_{ij} depends on $\{L_{jk} : \; (j, k) \in E, 1 \le k < j\}$, $\{L_{ik} : \; (i, k) \in E, 1 \le k < j\}$ and $\{D_{kk}, 1 \le k \le j\}$, hence it is clear that Ω_{ij} is functionally independent of elements of $\Theta_{\mathcal{G}}$ that follow it in the arrangement described above. Hence the gradient matrix of ψ (with this arrangement) is a lower triangular matrix, and the Jacobian of ψ is therefore given as

$$\prod_{i=1}^m \left(\prod_{j \in ch(i)} D_{jj}^{-1}\right)\prod_{i=1}^m D_{ii}^{-2}.$$

It follows from the expression above that the Jacobian of ψ is

$$\prod_{j=1}^m D_{jj}^{-(pa_j+2)}.$$

Proof of Proposition 8.1.

$$\begin{aligned}
\mathbb{E}[\Omega|D] = \mathbb{E}[LD^{-1}L^t|D] &= \mathbb{E}[\sum_{i=1}^{m} D_{ii}^{-1} L_{\cdot i} L_{\cdot i}^t | D] \\
&= \sum_{i=1}^{m} D_{ii}^{-1} \mathbb{E}[\begin{pmatrix} 1 \\ \beta_{<i]} \end{pmatrix} [1, \beta_{<i]}^t] \Big)^0 |D] \\
&= \sum_{i=1}^{m} D_{ii}^{-1} \mathbb{E}[\begin{pmatrix} 1 & \beta_{<i]}^t \\ \beta_{<i]} & \beta_{<i]}\beta_{<i]}^t \end{pmatrix}^0 |D] \\
&= \sum_{i=1}^{m} D_{ii}^{-1} \begin{pmatrix} 1 & -U_{[i>} U_{<i>}^{-1} \\ -U_{<i>}^{-1} U_{<i]} & \mathbb{E}[\beta_{<i]}\beta_{<i]}^t|D] \end{pmatrix}^0 . \qquad (12.8)
\end{aligned}$$

The conditional expectation in Equation (12.8) can be noted by computing the following:

$$\begin{aligned}
\mathbb{E}[\beta_{<i]}\beta_{<i]}^t|D] &= \mathbb{V}ar[\beta_{<i]}|D] + \mathbb{E}[\beta_{<i]}|D]\mathbb{E}[\beta_{<i]}|D]^t \\
&= D_{ii}U_{<i>}^{-1} + U_{<i>}^{-1}U_{<i]}U_{[i>}U_{<i>}^{-1} \qquad (12.9)
\end{aligned}$$

Now since $D_{ii}^{-1} \sim G(\alpha_i/2 - pa_i/2 - 1, 2U_{ii|<i>}^{-1})$, $\mathbb{E}[D_{ii}^{-1}] = (\alpha_i - pa_i - 2)U_{ii|<i>}^{-1}$ and therefore

$$\begin{aligned}
\mathbb{E}[\Omega] &= \sum_{i=1}^{m} \begin{pmatrix} (\alpha_i - pa_i - 2)U_{ii|<i>}^{-1} & (\alpha_i - pa_i - 2)(-U_{ii|<i>}^{-1}U_{[i>}U_{<i>}^{-1}) \\ (\alpha_i - pa_i - 2)(-U_{<i>}^{-1}U_{<i]}U_{ii|<i>}^{-1}) & U_{<i>}^{-1} + (\alpha_i - pa_i - 2)U_{<i>}^{-1}U_{<i]}U_{ii|<i>}^{-1}U_{[i>}U_{<i>}^{-1} \end{pmatrix}^0 \\
&= \sum_{i=1}^{m} (\alpha_i - pa_i - 2) \begin{pmatrix} U_{ii|<i>}^{-1} & -U_{ii|<i>}^{-1}U_{[i>}U_{<i>}^{-1} \\ -U_{<i>}^{-1}U_{<i]}U_{ii|<i>}^{-1} & U_{<i>}^{-1} + U_{<i>}^{-1}U_{<i]}U_{ii|<i>}^{-1}U_{[i>}U_{<i>}^{-1} \end{pmatrix}^0 \\
&\quad - \sum_{i=1}^{m} (\alpha_i - pa_i - 3) \left(U_{<i>}^{-1}\right)^0 \\
&= \sum_{i=1}^{m} (\alpha_i - pa_i - 2) \left(U_{\leq i\geq}^{-1}\right)^0 - \sum_{i=1}^{m} (\alpha_i - pa_i - 3) \left(U_{<i>}^{-1}\right)^0 .
\end{aligned}$$

Proof of Proposition 8.2.

First recall that from Equation (4.5) that for each $i \in V$

(i) $\Sigma_{ii} = \lambda_i + \beta_{[i>}^t \Sigma_{<i>} \beta_{<i]} = \lambda_i + \text{tr}(\Sigma_{<i>}\beta_{<i]}\beta_{[i>}^t)$ and

(ii) $\Sigma_{<i]} = \Sigma_{<i>}\beta_{<i]}$.

Starting from the largest index $m \in V$ we have $\Sigma_{[m>} = 0$ and $\Sigma_{mm} = \lambda_m$. Therefore $\mathbb{E}[\Sigma_{mm}] = \dfrac{U_{mm}}{\alpha_m - 4}$ as $\lambda_i \sim IG(\frac{\alpha_m}{2} - 1, \frac{1}{2}U_{ii|<i>})$. Now suppose that for some $1 < i < m$ the expected values of Σ_{kl} for all $k, l \in pa(i)$, has been calculated, i.e., $\mathbb{E}[\Sigma_{<i>}]$ is known. Using part (ii) above and the fact that $\Sigma_{<i>} \perp \Sigma_{<i>}^{-1}\Sigma_{<i]}$ due to the mutual independence property of $\{(D_{ii}, L_{<i]}) : i = 1, \ldots, m\}$ as given by Theorem 6.1 we obtain

$$\mathbb{E}[\Sigma_{<i]}] = -\mathbb{E}[\Sigma_{<i>}]U_{<i>}^{-1}U_{<i]}.$$

Now applying part (i), Equation (6.1) from Theorem 6.1 and Equation (12.9) we obtain

$$\begin{aligned}\mathbb{E}[\Sigma_{ii}] &= \mathbb{E}[\lambda_i] + \operatorname{tr}\left(\mathbb{E}[\Sigma_{<i>}\beta_{<i]}\beta_{[i>}^t]\right)\\ &= \frac{U_{ii|<i>}}{\alpha_i - pa_i - 4} + \operatorname{tr}\left(\mathbb{E}[\Sigma_{<i>}]\mathbb{E}\left[\mathbb{E}[\beta_{<i]}\beta_{<i]}^t|D_{ii}]\right]\right)\\ &= \frac{U_{ii|<i>}}{\alpha_i - pa_i - 4} + \operatorname{tr}\left(\mathbb{E}[\Sigma_{<i>}]\left(\frac{U_{ii|<i>}U_{<i>}^{-1}}{\alpha_i - pa_i - 4} + U_{<i>}^{-1}U_{<i]}U_{[i>}U_{<i>}^{-1}\right)\right).\end{aligned}$$

13 Hausdorff measures

In order to derive the density of $\pi_{U,\alpha}^{P_{\mathcal{G}}}$ we begin with a short introduction to Hausdorff measures. The reader is referred to [2, Section 19] for more details on this topic.

Let $\mathscr{X} = (\mathcal{X}, d)$ be a metric space, δ a non-negative real number and V a subset of $\mathcal{X}$. A δ-cover of V is a finite or infinite sequence $\{U_i, i \in I\}$ of subsets of $\mathcal{X}$ such that

$$V \subset \cup\{U_i : i \in I\} \quad \text{and} \quad d(U_i) = \sup\{d(x,y) : x, y \in U_i\} < \delta, \quad \forall i \in I.$$

Given $r > 0$ we define a set function

$$\mathcal{H}_{\delta,\mathscr{X}}^r(V) := \inf\left\{\sum_{i\in I} d(U_i)^r : \{U_i : i \in I\} \text{ is a } \delta\text{-cover of } V\right\}. \tag{13.1}$$

Note that the infimum is taken over all δ-covers of V. If no such cover exists, then the infimum is $+\infty$. The r-dimensional Hausdorff (outer) measure $\mathcal{H}_{\mathscr{X}}^r$ is now defined as follows.

$$\mathcal{H}_{\mathscr{X}}^r(V) = c_r \lim_{\delta\to 0} \mathcal{H}_{\delta,\mathscr{X}}^r(V),$$

where c_r is an optional normalizing constant. From Equation (13.1) it is clear that when V is a subset of $\mathcal{X}_0 \subset \mathcal{X}$ it is enough to include δ-covers consisting of the subsets of $\mathcal{X}_0$ alone. In this framework, when $\mathscr{X}$ is the Euclidean space $(\mathbb{R}^n, \|\cdot\|)$, without raising any ambiguity, we suppress the under-script $\mathscr{X}$ and write $\mathcal{H}^r$ for the r-dimensional Hausdorff measure on $\mathbb{R}^n$. Furthermore, in this case we choose the normalizing constant c_r to be the volume of the r-dimensional ball of diameter 1 in $\mathbb{R}^r$. By incorporating this normalizing constant, $\mathcal{H}^r$ coincides with the standard Lebesgue measure on $\mathbb{R}^r$.

13.1 Integration and change of variable

We now proceed to discuss integration and change of variable in the context of Hausdorff measures. For a $k \times n$ matrix $A \in \mathbb{R}^{k\times n}$ let us define $|A| = \sqrt{\det(A^tA)}$. More generally, if $T : \mathbb{R}^n \to \mathbb{R}^k$ is a linear mapping, then we define $|T|$ in terms, but clearly independent, of a matrix representation of T. If T is one-to-one, then $\mathcal{H}^r(T(V)) = |T|\ \mathcal{H}^r(V)$, for each $V \subset \mathbb{R}^n$.

Now suppose f is a continuous, one-to-one mapping from an open subset V of $\mathbb{R}^n$ into $\mathbb{R}^k$. If f has continuous partial derivatives, then by Theorem 19.3 in [2] we have:

$$\int_V g(f(x))\mathrm{J}(f(x))\lambda^n(dx) = \int_{f(V)} g(y)\mathcal{H}^n(dy), \tag{13.2}$$

where J($f(x)$) is the (Hausdorff) Jacobian of f defined by $|Df(x)|$ and λ^n is the standard Lebesgue measure on $\mathbb{R}^n$.